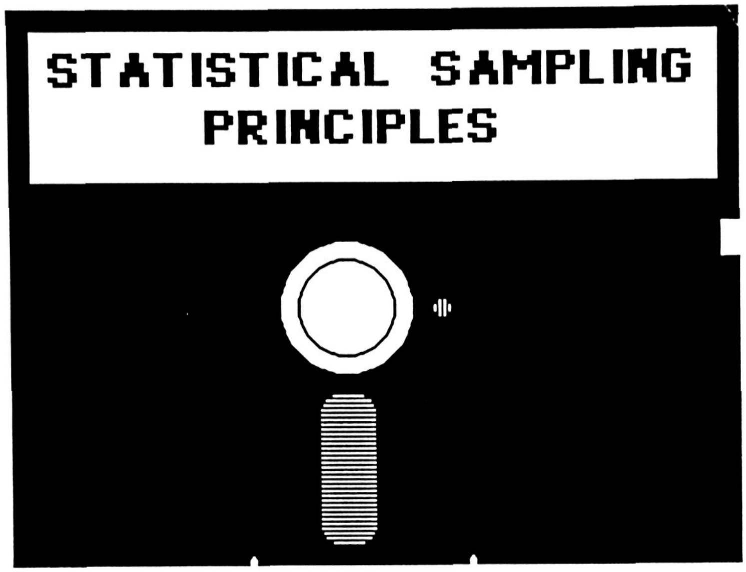
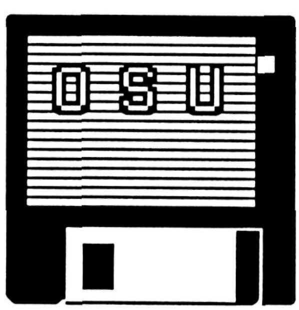

STATISTICAL SAMPLING
PRINCIPLES


OSU


HERBERT W.
OCKERMAN

STATISTICAL SAMPLING PRINCIPLES

by
HERBERT W. OCKERMAN
Department of Animal Science
The Ohio State University
and
The Ohio Agricultural
Research and Development Center

INDEX

INDEX (Continued)

INDEX (Continued)

INTRODUCTION

In the not too distant past, reading, writing and arithmetic were the foundation necessary for good citizenship and a successful business life. As we pass from the atomic age into what is known as the computer or statistical age the arithmetic cornerstone of our foundation is expanding and a knowledge of the methods and terminology of the statistical branch of applied mathematics is necessary for effective communication and even to comprehend the daily written word.

Statistical terms and measurements bombard us every day and a number of them have a very serious effect on our lives. In fact, most decisions are essentially based on information that is statistical in nature. An example of a few of these are listed as follows:

1. Time of sunrise tomorrow
2. Average temperature (also maximum and minimum temperatures) for this day of the year
3. Odds on rainfall in the next 24 hours
4. Cost of living index
5. Gross national product
6. Change in value of money
7. Election or opinion polls and early prediction of winners
8. Estimated population of the U. S.
9. Mortality rates used to adjust insurance premiums
10. Percentage of unemployment
11. Relationship between cigarette smoking and cancer
12. Rate of development of the "average" child
13. Children's clothing sizes (3-year old)

14. Livestock market prices
15. Stock market averages
16. Prediction of total wheat that will be harvested next year
17. Average consumption of a given product
18. TV program ratings
19. Batting average in baseball
20. Gambling odds or risk
21. Profit or loss charts

Statistics, in general, is the science of dealing with numerical data. It supplies the tools for collecting, analyzing, interpreting and projecting these data. Variable individual events if observed as a group will normally yield trends. Statistics may be defined as the tools necessary to obtain group trends by collecting and objectively evaluating separate individual events and then projecting how this trend will behave (within limits). In simple terms, statistics is a mathematical tool for predicting under a given set of conditions what will happen (also tells the odds of the prediction being wrong). Like all tools, if placed in the hands of a craftsman, they can be extremely helpful. If, however, they are misused (by ignorance or design) the end results can be useless and, in many cases, misleading and harmful.

Statistical tools are mathematical in nature but statistical application and interpretation requires logic as well as mathematical skills. Just as you can add the number of oranges and apples mathematically you can perform statistical problems that are incorrect logically and obtain an answer (which is worthless).

Even on a very basic level it is necessary to understand such things as:

1. Sampling
2. Averages
3. Variation
4. Relationship
5. Understanding charts and tables

People who do not understand statistics usually react to statistics in one of the following extreme manners:

1. Accept all conclusions obtained from statistical mathematics regardless of how illogically they were obtained or how fantastic they appear.
2. Extreme scepticism of all statistically derived information.

Both of these extreme viewpoints indicate the lack of statistical understanding and deprive the individual of a valuable tool, to aid in his decision-making task. A person knowledgeable in statistical methods will question the logic behind the experiment and the analysis and take into consideration the possible bias. If satisfied that everything is in order he will then accept the results and know the odds of having made an incorrect judgment.

Today Quality Control budgets in the Meat Area average from 1-1/2 to 3% of total sales.

A beginning student often wonders why samples are necessary--why not observe the total group and secure a greater degree of accuracy. The answer to this question usually involves time and money and these two items become a limiting factor in all societies. Even the assumption that the accuracy would be improved may be in error particularly if time is limited (it always is) or the product is changing (time is again limited). A few accurately taken samples are more useful than a hurried total analysis. Often the sampling procedure destroys the product (nitrogen analysis of meat) and consequently if the total product is analyzed nothing is left.

The basic steps in statistically analyzing a problem would include:

1. Define the problem (what questions are to be answered).
2. What is the size and nature of the group to which these answers are to apply?
3. Designing an experiment to answer the problem which include:
    a. methods of selecting the sample (eliminate bias)
    b. determining what observations are necessary to answer the question and how these observations are to be made.
4. Collecting the data
5. Assembling or classifying the data
6. Summarizing the data. (Usually the statistical calculation step)
7. Presenting the results
    a. table or graph form
    b. projection of future events (inference)
8. Decision made on results

As can be seen many steps are involved in statistical analysis in addition to the mathematical manipulation and like a chain, decisions based on these results are no stronger than the weakest link.

Statistics is often classified into two categories and actually these are a summary of the above steps.

1. Descriptive - This portion of statistics describes the sample observed (Steps 3 through 7a)
2. Inductive - This portion of statistics generalizes, predicts, or estimates from a sample to the total group. (Step 7b)

HOW WELL DOES 100% INSPECTION WORK?

Most people assume that 100% inspection will give complete assurance that inspected lots will be effectively separated.

The following are 450 random three digit numbers. Assume that the numbers in the range 621 through 874 inclusive are defectives. Proceed through this "lot" once and see how effectively you can spot and check each defective sample you find. Allow a maximum of six minutes to complete the inspection. Count the total number of defectives you found.

570	209	201	362	607	647	819	917	480	501
666	340	559	936	995	393	850	919	368	657
680	320	252	493	440	265	974	480	130	836
849	570	624	713	241	680	612	637	167	309
110	606	183	496	984	656	167	178	570	997
866	763	061	277	052	988	161	188	656	886
916	150	100	906	855	751	422	349	921	690
213	488	642	440	813	994	699	101	562	290
425	600	546	690	714	972	071	092	301	197
678	125	366	253	031	281	107	810	579	434
439	179	334	081	558	988	539	953	475	685
381	078	174	530	458	709	465	496	066	675
547	315	823	901	406	276	336	993	918	967
234	208	823	521	780	427	033	944	553	096
115	082	263	2 Min. → 300	415	737	923	348	429	396
104	263	752	015	211	689	853	226	365	112
372	859	104	977	835	178	082	391	119	733
105	889	482	772	671	774	698	893	966	964
960	539	182	494	242	259	511	630	085	276
835	652	240	011	888	268	608	988	368	138
132	850	336	715	085	904	944	414	011	409
738	643	636	234	995	586	582	451	162	391
916	461	378	518	851	998	099	669	056	762
331	312	232	426	958	172	174	070	736	509
421	221	582	591	493	346	140	016	663	124
524	066	242	581	751	103	740	530	164	849
012	069	559	358	506	200	248	800	639	236
367	411	763	465	357	308	874	025	334	700
586	676	883	500	101	351	072	242	488	306
837	363	699	731	162	731	917	878	115	315
300	147	483	767	397	423	4 Min. → 963	265	525	229
259	126	808	866	845	432	263	085	676	059
760	960	906	989	947	432	664	002	742	739
470	661	566	457	466	422	266	404	366	421
402	645	411	715	500	305	941	594	921	641
916	901	673	324	005	206	417	766	172	481
808	426	791	818	317	341	772	512	582	471
038	931	279	846	537	170	565	807	725	988
841	599	587	408	501	293	557	735	579	011
611	697	988	654	324	604	889	870	794	976
952	416	357	059	198	908	127	831	347	763
309	988	917	883	972	024	944	044	837	593
080	848	795	555	836	134	301	638	259	567
708	020	099	180	659	127	495	993	634	6 Min. 955
942	154	884	281	668	618	492	103	858	503

If 100 percent inspection is the answer to total accuracy then all students should have obtained the same to the preceding problem. Normally, this does not occur and the reason for the differences in results is the limiting factor of time.

References

Chou, Ya-Lun. 1965. Applied Business and Economic Statistics. Holt, Rinehart and Winston, New York, Chicago, San Francisco, Toronto, London.

McCarthy, Philip J. 1957. Introduction to Statistical Reasoning. McGraw-Hill Book Company, Inc., New York, Toronto, London.

Walker, Helen M. 1958. Elementary Statistical Methods. Holt, Rinehart, and Winston, New York.

POPULATION VS. SAMPLE

In most statistical problems we are interested in estimating values for a population.

A population may be defined as the complete or entire group for which estimating values are needed. Thus, the population is the group we seek knowledge about but in most practical statistical problems, these values are not known and only estimates of the values will be obtained (not the actual values themselves). Examples of populations are as follows:

1. Hams produced by Company A last year
2. Hams produced by Company A on January 21 of last year
3. Hams produced by Company A in Plant B on January 21 of last year
4. Hams produced by Company A in Plant B on January 21 of last year on the 4th smoking period of smokehouse (i).

The above list of examples illustrate the necessity of being very specific when describing a population so that the selected samples may represent this population and the results of these samples may be applied only within the boundaries set by this population. Therefore, there should be no confusion on the part of the individual taking the samples and the individual using the results as to the exact boundaries of the population involved.

Since populations are often large, the necessity of sampling this total is usually encountered.

A sample is defined as a subdivision or portion of the population that is used to obtain information on the population. We observe the sample in an attempt to gain an estimate of facts about the total population. An example of samples is as follows:

1. Two hams, per plant, per week for Company A during last year.
2. Ten hams, per plant, produced on January 21 of last year for Company A.
3. Fifty hams produced in Plant B for Company A on January 21 of last year.
4. 10% of the hams smoked in smokehouse (i) on the 4th smoking period in Plant B for Company A on January 21 of last year.

As is illustrated in the above examples when the population changes the sampling procedure normally will also change.

After the sample has been taken it is necessary to evaluate it. This evaluation can result in two types of values as follows:

1. Constant - This characteristic value cannot vary. Example would be the number of femur bones in a whole ham.
2. Variable - A characteristic of an individual sample that may have different values.

   There are two types of variables that are encountered in statistical work and they are as follows:

   a. Continuous variable - This type of variable may assume any value between given limits. The measurement types of evaluations would be included in this type of variable.

      Example - weight of a ham.

      (1) 14 lbs. 2 oz.

      (2) 16.34 pounds

      (3) 12.3496 pounds

      (4) 6321.6432 grams

      The only reason a ham could not have a specific weight (between maximum and minimum limits) is the lack of accuracy of the weighing device.

b. Discrete variable (often called an attribute) - This type of variable may only assume specific values between given limits. The counting types (can only be whole numbers) of evaluations would be included in this type of variable.
Examples:

(1) Number of whole hams placed in a smokehouse

(a) 48 hams

(b) Cannot be 48.67

(2) Number of "heads" received in 50 tosses of a coin

Another example of a discrete variable would be the following classification of defects:

(1) Critical (2) Major (3) Minor

Defects must fall into a specific category and cannot take intermediate values.

After the variable to be evaluated has been determined, it is then necessary to collect data on this variable.

Data may be defined as the values a variable assumes. Examples of data values would be as follows:

Ham weight

(1) 16.3 lbs. (2) 14.6 lbs

Percent of unacceptable products

(1) 5.0% (2) 4.3%

Wieners per smokehouse stick

(1) 160 (2) 168

After the data have been collected and classified a statistic is calculated for the sample. (English letters are used as abbreviations for the statistical values of a sample.)

A <u>statistic</u> may be defined as information gained from a sample that is used to estimate population information. An example would be:

Sample average - - $\bar{X}$ = total data/number of samples

A <u>parameter</u> may be defined as a characteristic of a population. (Greek letters will be used as abbreviations for the parameter values of a population). These values are usually never known and are estimated by the sample statistic.

An example would be: population average (average weight of wieners produced in a plant in a day). The Greek letter μ is used for this parameter abbreviation.

An example of most of the previously defined terms may be illustrated in the following examples:

1. Population - Canned ham production of Plant A on the 6th day of February of this year.
2. Sample - 2 cans from each retort cooking on the 6th day of February of this year in Plant A.
3. Variable -
   a. continuous - weight of can (will be used in the remainder of this example)
   b. discrete - number of dented cans (will not be used in remainder of example)
4. Data - Actual weight of the individual cans in the sample.
5. Statistic - Average weight of the cans in the sample. ($\bar{X}$)
6. Parameter - Average weight of all the cans produced in Plant A on the 6th day of February of this year (μ). This value will never be known since all the cans were not weighed but may be estimated from the sample average. ($\bar{X}$)

A control chart is a graph on which a center horizontal line and an upper and lower horizontal limits are plotted for a simple statistic. Data for this same statistic is then plotted for a series of samples.

The <u>center line</u> on the control chart represents the expected value (aiming point) of the statistic to be plotted.

The <u>control limits</u> on the control chart are used as guidelines for judging the significance of the variation of the statistical measures plotted for sub-groups or individuals.

References

Haber, Audrey and Richard P. Runyon. 1969. General Statistics. Addison-Wesley Publishing Company., Reading, Massachusetts; Menlo Park, California; London; Don Mills, Ontario.

Snedecor, George W. 1956. Statistical Methods. The Iowa State College Press, Ames, Iowa.

Spiegel, Murray R. 1961. Theory and Problems of Statistics. McGraw-Hill Book Company, New York, St. Louis, San Francisco, Toronto, Sydney.

ORGANIZING AND DISPLAYING DATA

Most data are collected and recorded in a raw data table. This type of table is usually organized so that data may be rapidly located but the table size is usually so large that summarization and trends are difficult to observe. For this purpose a summary table is usually constructed from the raw data table and is a reorganization and condensation of the original data. The definition of a table is the systematic arrangement of data into columns and rows. The procedure normally used is to separate the observations into groups and report the totals, averages and percentages of these groups. The following table is an example of this type of summary.

Cash Receipts from Agricultural Marketing in Ohio[1]/

Year	Livestock	Crops	Total	Livestock	Crops
	Thousand Dollars			Percent	
1930	234,402	80,934	315,336	74.3	25.7
1940	225,997	91,622	317,619	71.2	28.8
1950	607,320	287,640	894,960	67.9	32.1
1960	602,906	402,077	1,004,983	60.0	40.0
1970	778,106	599,739	1,377,845	56.5	43.5
1980	1,380,138	2,801,733	4,181,871	33.0	67.0

1/ OARDC, 1968-78, 82; "1967-77,81 - Ohio Farm Income", Wooster, Ohio.

This table separates the cash receipts from Ohio Agricultural Marketing into time periods of one year each. It reports cash receipts separately for livestock and crops in thousand dollar units. It also reports the percentage of cash receipts derived from livestock and crops in Ohio for these same time periods.

Percentages are one of the most misunderstood and often misused forms of simple statistics. For percentages to have any meaning the base on which they are calculated must be clearly understood. For example, if a 100-pound meat product has the following composition:

70 pounds of water
18 pounds of protein
1 pound of ash
11 pounds of fat

The percent fat based on the wet sample would be:

$$\frac{11}{100} \times 100 = 11\%$$

The percent fat based on a dry sample would be:

$$\frac{11}{30} \times 100 = 36.6\%$$

The quantity of fat has not changed but the percentages are drastically different because different bases were used.

Another example would be a 14-pound fresh (green) ham pumped with 1.4 pounds of cure. The percentage increase in weight would be:

Based on green weight -

$$\frac{1.4}{14} \times 100 = 10\% \text{ increase in weight}$$

Based on final weight -

$$\frac{1.4}{14+1.4} \times 100 = 9.1\%$$

Again the increase in weight was the same but the base and the calculated percent were different.

Other examples in the meat processing area would include the following type of questions that concern the base of the percentage calculation:

1. Percent breading on a product - is the base, of the percentage, green weight or final weight?
2. Percentage pump on a boned ham - is the base, of the percentage, bone-in weight or boned weight?

Another property of percentages is that the total component percentages must equal 100. If only one quantity is increased (or decreased) it will decrease (or increase) the percentage value of the other components. Again we will return to our original 100 pounds of meat for an example:

Original Sample		To this sample	New Sample	
70 pounds water	(70%)	is added	70 pounds of water	(56.0%)
18 pounds of protein	(18%)	25 pounds	18 Pounds of protein	(14.4%)
1 pound of ash	(1%)	of fat	1 pound of ash	(.8%)
11 pounds of fat	(11%)	⟶	36 pounds of fat	(28.8%)
100 " of product	(100%)		125 " of product	(100.0%)

By addition of only fat to the original sample you will notice that the percentage of fat increased and also the percentage of all other components (water, protein and ash) decreased.

The converse is also true as will be shown by heating the original sample.

Original Sample				
70 pounds water	(70%)	⟶	45 pounds of water	(60.0%)
18 pounds protein	(18%)	loss of 25 pounds	18 pounds of protein	(24.0%)
1 pound of ash	(1%)	of water due to	1 pound of ash	(1.3%)
11 pounds of fat	(11%)	heating	11 pounds of fat	(14.7%)
100 " of product	(100%)		75 " of product	(100.0%)

In this case the decrease in water lowers the percent moisture and raises the percentage value of the other components (protein, ash and fat).

Often the first step in organizing table data is to group the observations into classes and construct a frequency distribution. A frequency distribution is defined as an arrangement of the data to show the frequency of occurrence of the observations in each of a small number of classes. An example of a frequency distribution table is as follows:

Farm Cash Receipts per County in Ohio in 1967[1/]

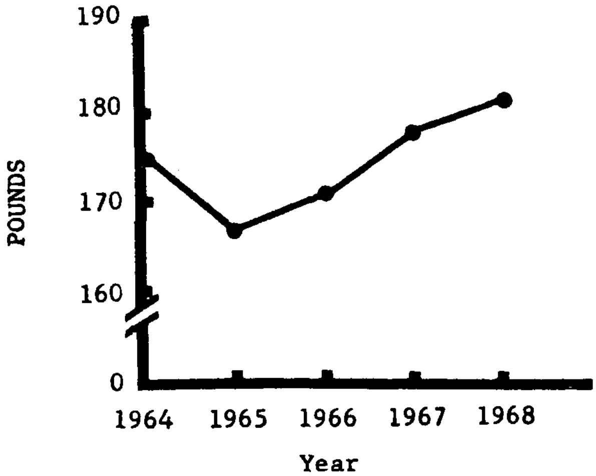
Total Cash Receipts (in Million Dollars)	Number of Counties	Number of Counties f
Greater than $20	卌 卌 卌 1111	19
$15 to $19.9	卌 卌 卌 卌 111	23
$10 to $14.9	卌 卌 卌 11	17
$5 to $9.9	卌 卌 卌 1	16
less than $5	卌 卌 111	13
		88

1/ OARDC, 1968; "1967 Ohio Farm Income", Wooster, Ohio

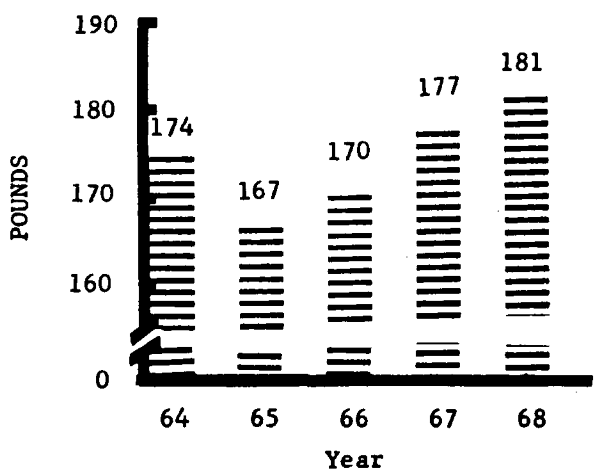
Determination of class intervals is an important step in setting up a frequency distribution table. The limits of each class must be clearly defined and the classes must be mutually exclusive so that it is impossible for an observation to fall into more than one class. The reason for grouping is to summarize the information and if the class intervals are extremely small, little summarization will take place. On the other hand, if the classes are extremely large most of the original information will have been lost. Therefore, the class size becomes a matter of judgment or compromise and the use for which the data are intended often will aid in this judgment.

Numerical data are often shown in visual form and this is referred to as a graph. A graph may be defined as a representation of numbers by geometric figures drawn to scale. Graphs may be of a number of types and a few examples will be shown on the following pages. The type of graph used will depend on the data, the audience for which it is intended, the points that are trying to be conveyed, ease of construction and the judgment of the person constructing this graph. The same information is shown on the following 3 types of graphs.

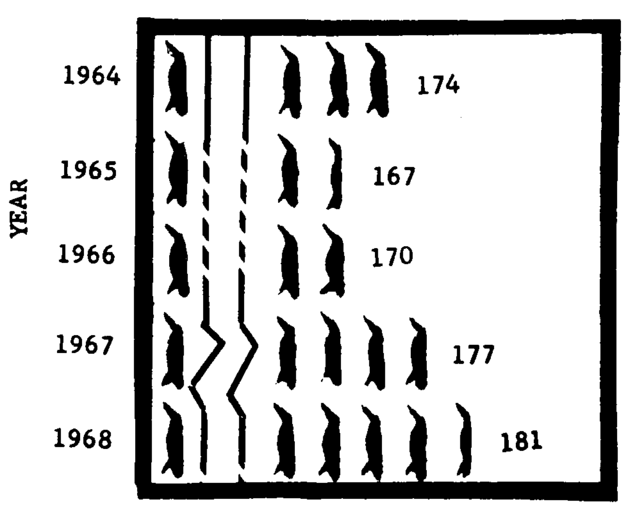
Per Capita Meat Consumption in U. S. (Carcass Wt.)



Line Graph



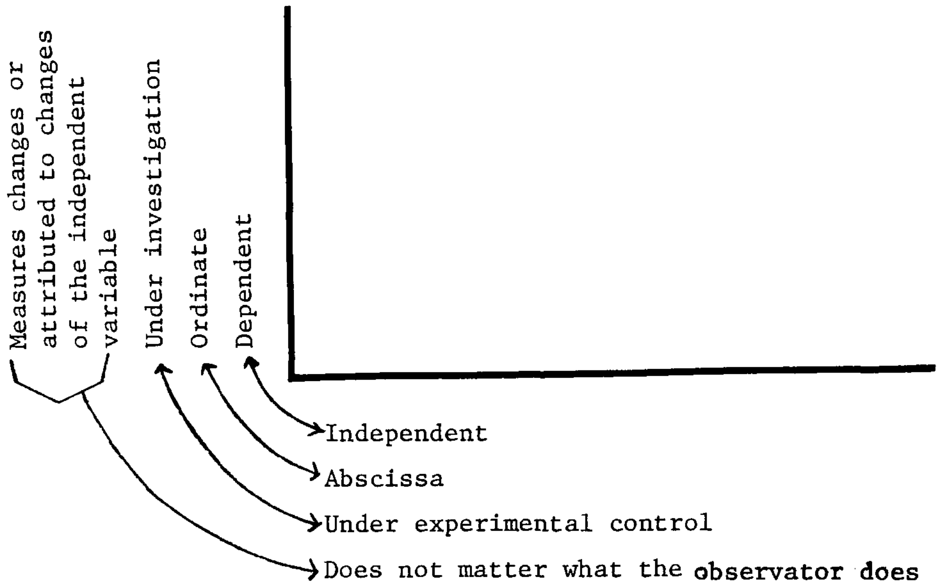
Bar Graph



Pictograph
Each Figure Represents 5 Pounds

USDA, 1968; National Food Situation, Nov. Issue.

In the line graph the dependent variable (pounds) is plotted on the vertical axis and the independent variable (years) is plotted on the horizontal axis.



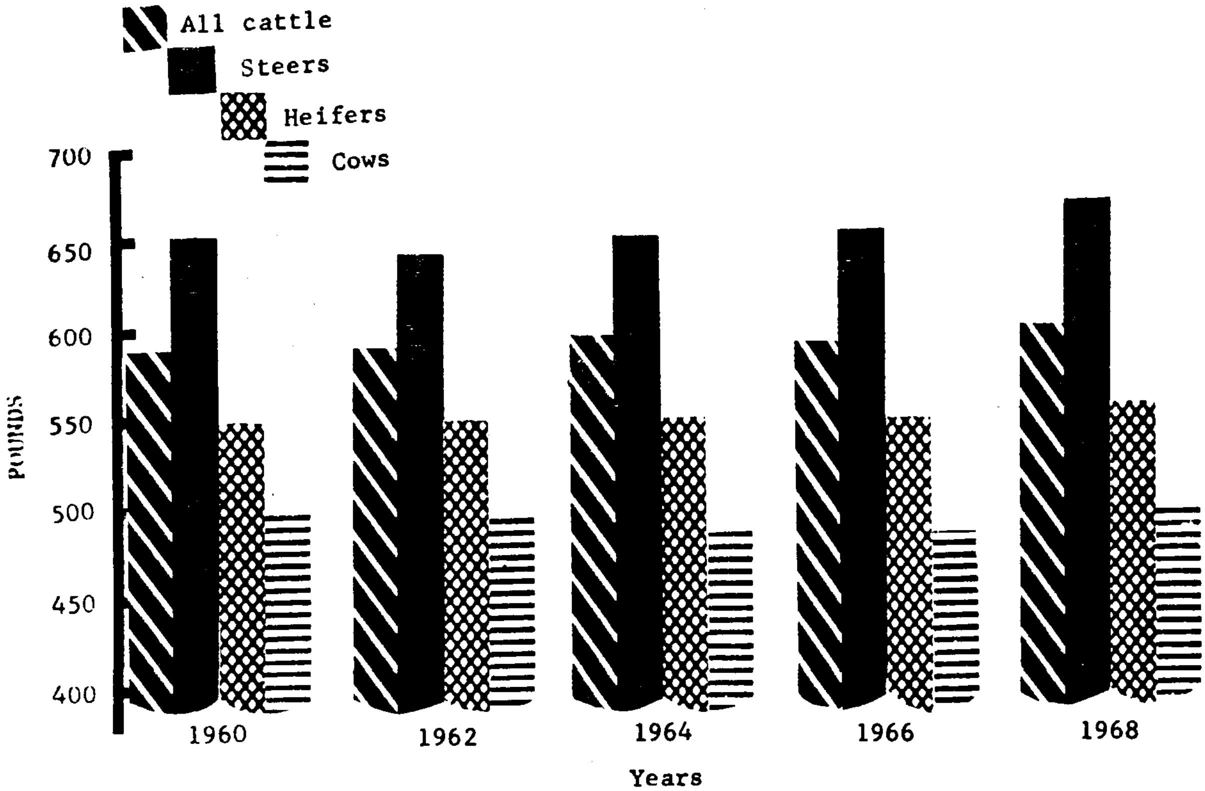
The scales on the two axes are not equal since they represent different types of measurements. If at all possible the vertical scale should include zero and if this is impractical (as in the example) the vertical scale should be broken to indicate the missing portion of the scale. The data on the vertical scale represent the average consumption for a calendar year. If data had been available only for the even number of years (1964 and 1966), these dots would have been connected with a straight line indicating that no information was available for 1965 (the line would have been above the actual 1965 level). Therefore, the straight lines between the dots indicate that no information is available in these areas.

The bar graph is similar to the line graph with the height of the bar representing the vertical quantity. The bar graph is called a histogram when the upper limits of one class are the lower limits of the next class and the vertical bars are adjacent to each other on the graph (not shown in example).

The pictograph is normally used when information is intended for the general public.

A graph showing several different types of things corresponding to one point on the horizontal axis may be illustrated by a line graph with more than one line or a bar graph with different types of bars. The following illustration is of this type of bar graph.

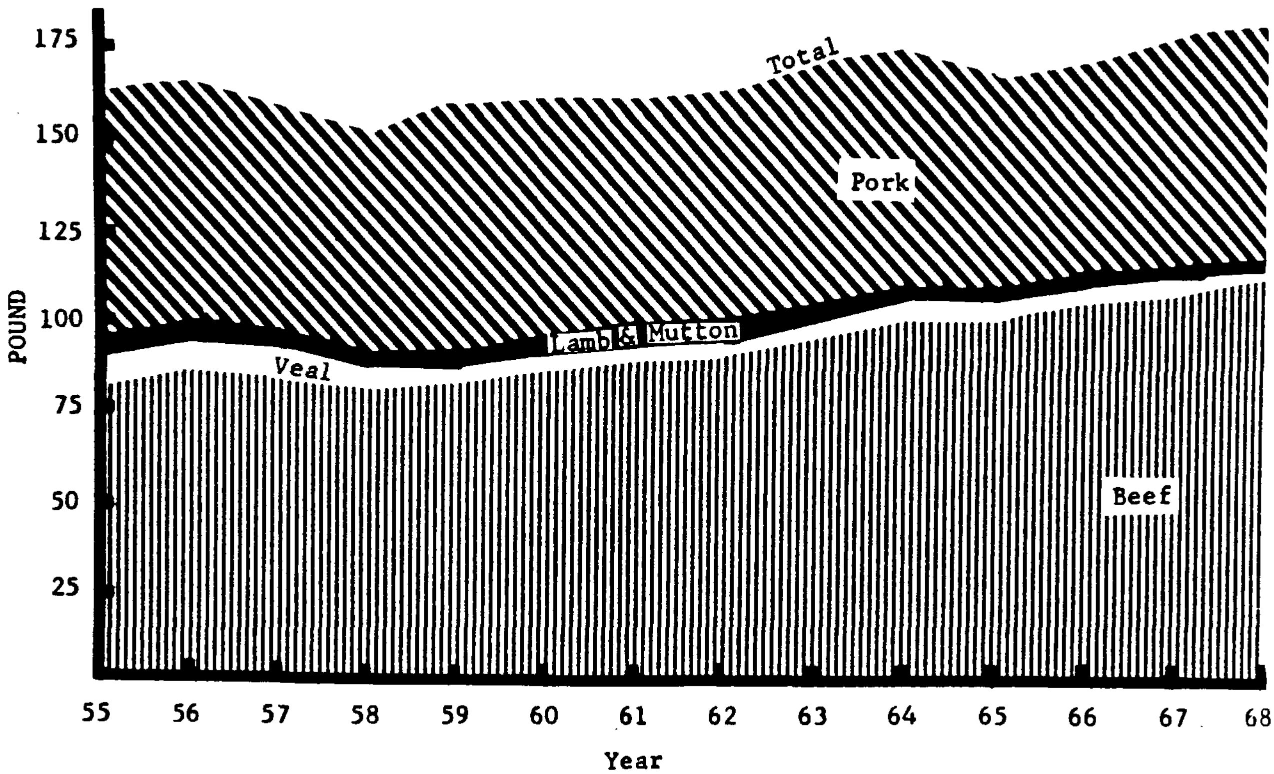
Average Dressed Cattle Weight. (under Federal Inspection)



USDA, 1968; Livestock & Meat Situation, Nov. Issue.

If the bars had been stacked (not overlapped), one upon the other for each year, this graph would have been called a <u>component part bar graph</u>. A component part line graph is illustrated in the next example. In this example the consumption of beef was plotted first. Then the quantity of veal consumed was plotted above the beef curve. Lamb and mutton, and pork were handled in the same manner. The top line on this graph gives the total consumption of red meat and the other line breaks this consumption down into its four major components.

Average Red Meat Consumption per Person in U.S. Divided Into Quantities of: Beef, Veal, Lamb & Mutton, and Pork



USDA, 1968; Livestock & Meat Situation. Nov. Issue.

On all graphs it is particularly important to pay close attention to the scales involved. By compressing or extending either the vertical or the horizontal scales it is possible to drastically change the visual appearance of the graph and to give the reader different impressions using the same data. In the following three graphs the same data are plotted but the scales have been changed.

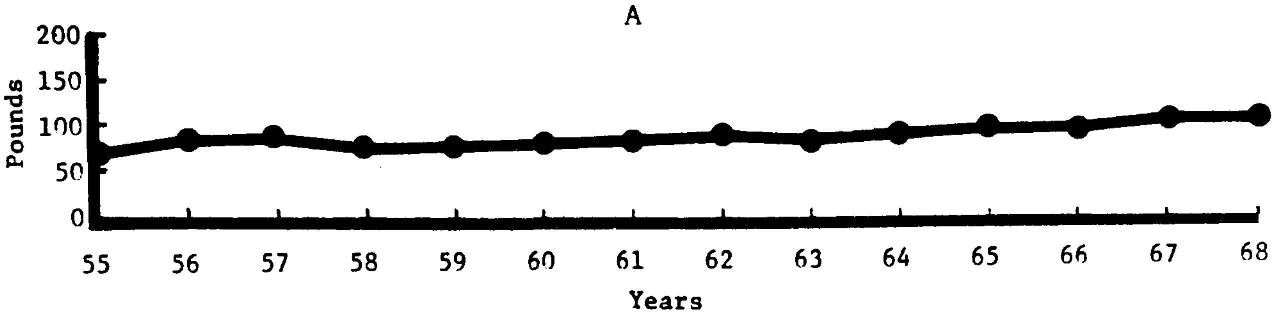
In graph A the vertical scale is compressed and the beef consumption appears to be fairly constant.

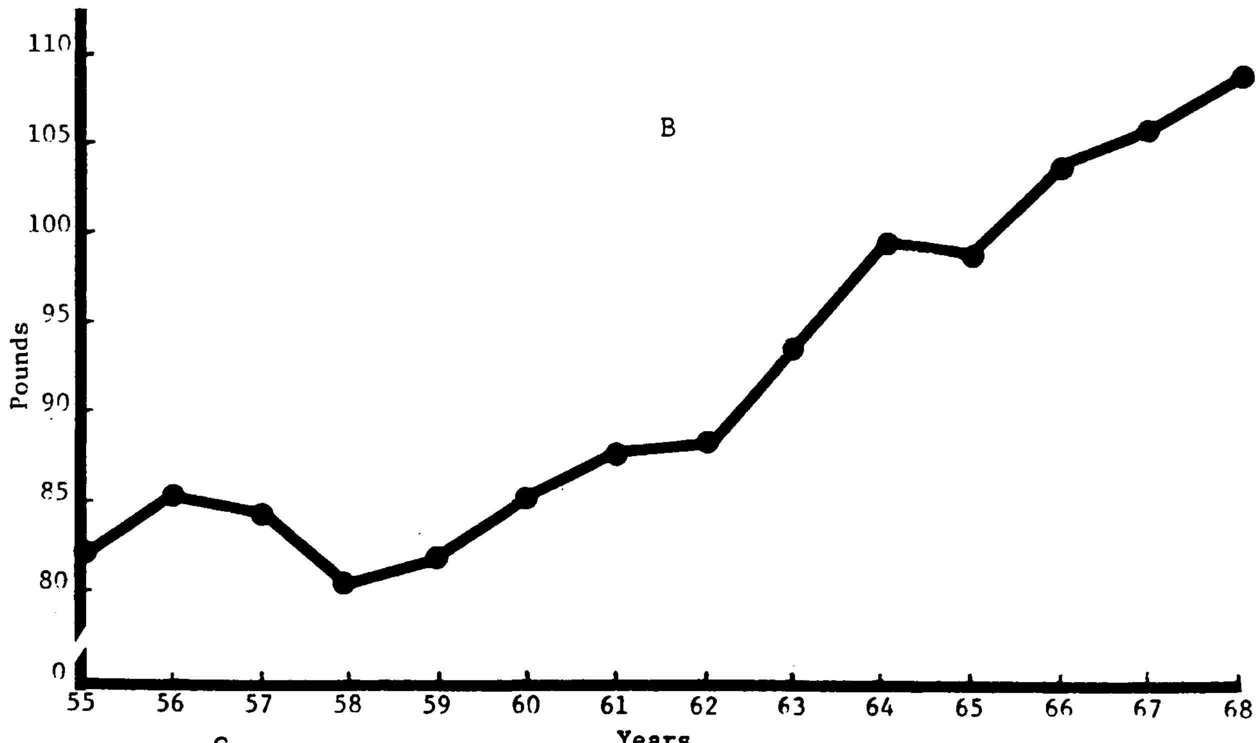
In graph B the vertical scale has been extended and beef consumption appears to be increasing at a moderate rate.

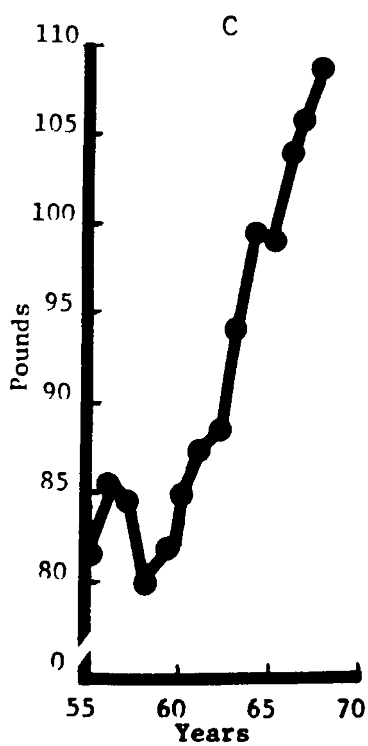
In graph C the vertical scale remains the same as in graph B but the horizontal scale has been compressed and the beef consumption appears to be rapidly advancing.

Remember, the data are the same, only the scales have been changed.

Beef Consumption Per Person







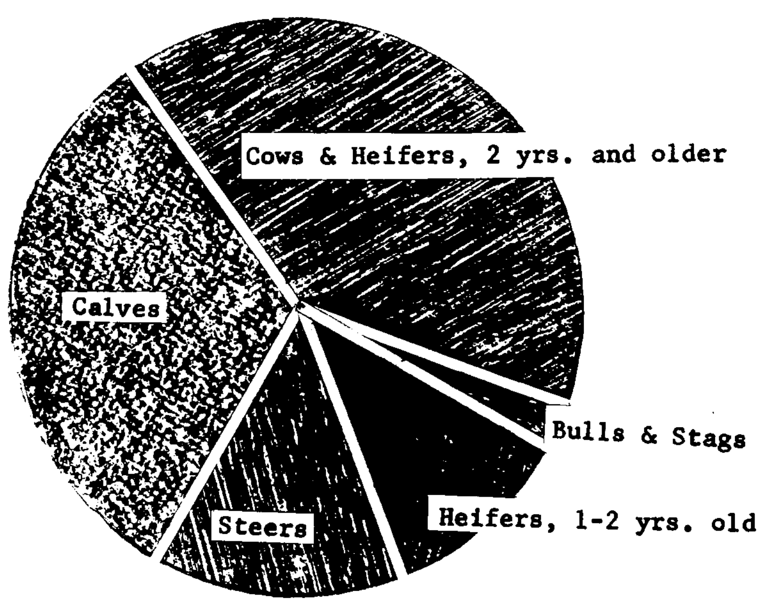
USDA, 1968: "Livestock & Meat Situation," Nov. Issue.

When it is necessary to show percentage type data, a pie chart or a percentage component part graph is often used. These graphs set the total area as 100 and then subdivide the component parts into proportional areas.

The following 2 pie charts show the beef cattle inventory in the U. S. in 1968 and placement of agriculture graduates in 1974. This type of chart gives no information as to numbers--only rations of the various components are shown.

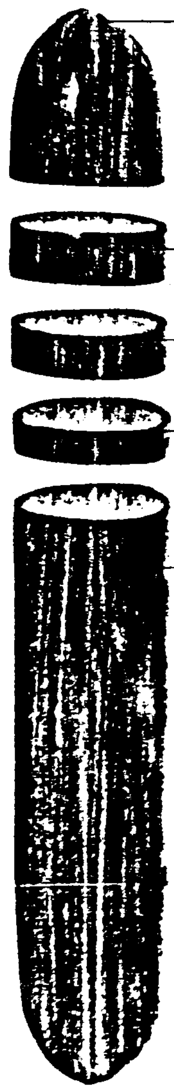
The next chart is the same type but is called a percentage component part bar graph. Again, ratio and not actual values are illustrated. For example, Fulton County, had $2,754,000 cash receipts from dairy products and this was 13.8% of its total agricultural cash receipts and Brown County had $2,336,000 from dairy products (less money than Fulton) but this was a larger percentage (37.9%) of its total agricultural cash receipts and this is the value shown on the bar graph.

Beef Cattle Inventory on Farms in U. S. on Jan. 1, 1968

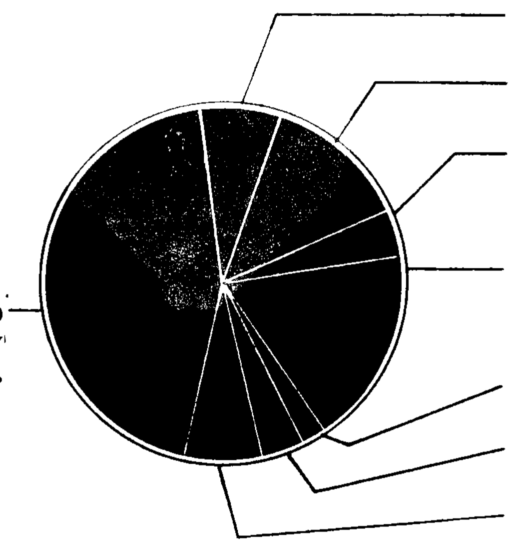


USDA, 1968; "Livestock & Meat Situation", Nov. Issue.

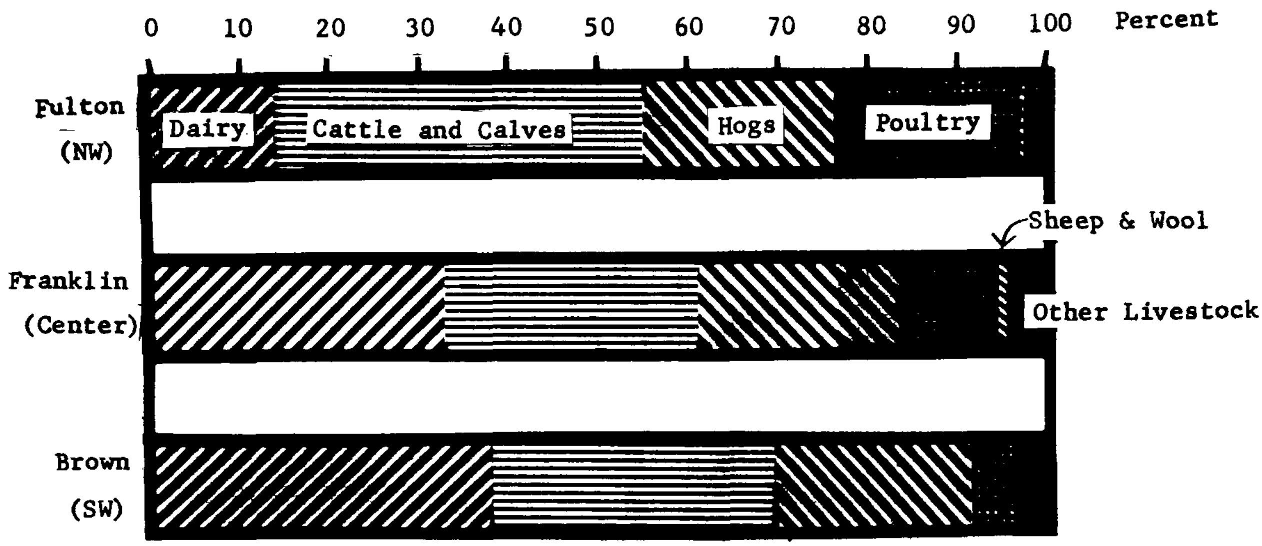
SALES DOLLARS



PLACEMENT OF 518 AGRICULTURE GRADUATES - 1975



Distribution of Receipts from Livestock & Livestock
Products for 3 Ohio Counties in 1967



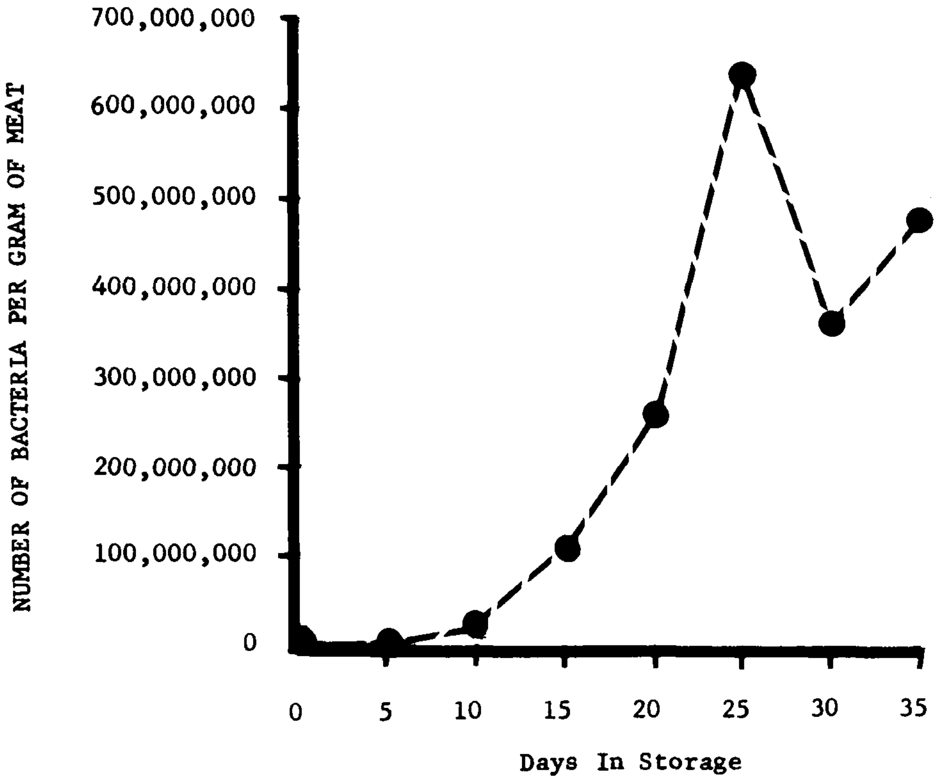
OARDC, 1968; "1967 - Ohio Farm Income". Wooster, Ohio

Scales other than arithmetical (same distance equals same quantity) are sometimes used on either of the two axes. The most popular is the use of the logarithmic scale (equal distance does not equal same quantity but equals a constant rate of change) on the vertical axes. This scale is used when ratio or relative amount of change is more important than the actual quantity of change. On this scale a straight line means a constant rate of change and the slope of the line indicates how rapid this change is occurring.

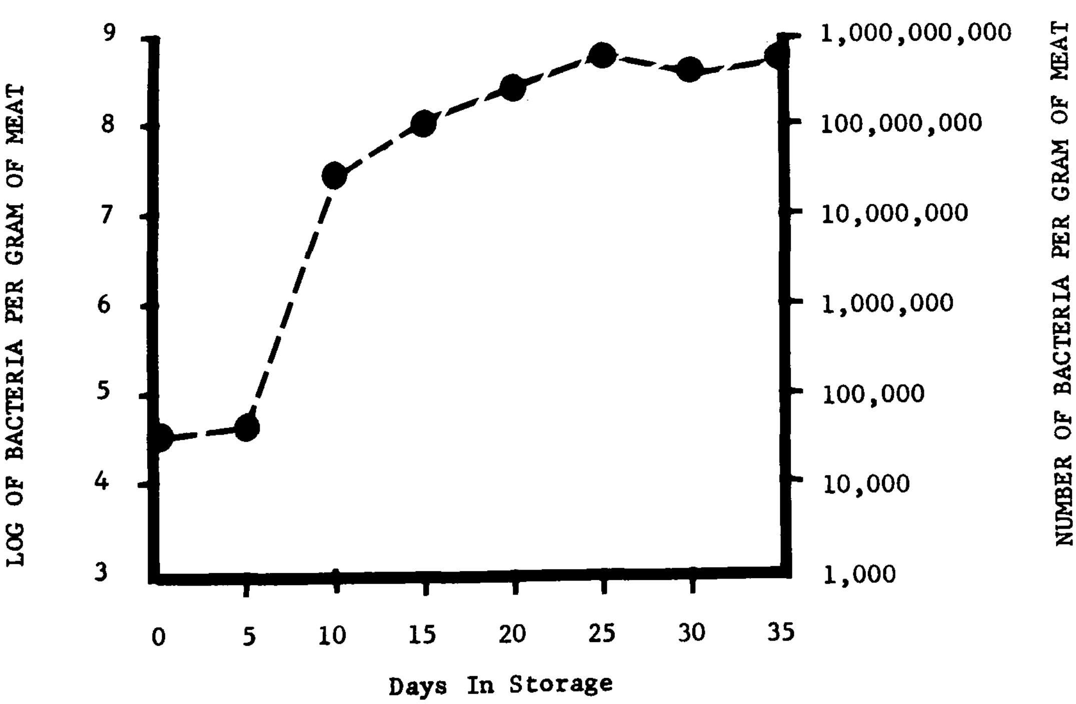
The next 2 graphs show the same microbiological data plotted on both types of scales. The impression given is quite different.

Number of bacteria per gram in a meat product

A. Plotted on an arithmetic scale



B. Plotted on a logarithmic scale



Tables and graphs have several uses in the statistical area and a few of these may be summarized as follows:

1. To summarize data.
2. To present data that will receive no further statistical treatment.
3. To summarize data that will subsequently be statistically analyzed and this summary will aid in checking the validity of the statistical calculations.

References

Chou, Ya-Lun. 1963. Applied Business and Economic Statistics. Holt, Rinehart, and Winston, New York, Chicago, San Francisco, Toronto, London.

Haber, Audrey and Richard P. Runyon. 1969. General Statistics. Addison-Wesley Publishing Company, Reading, Massachusetts; Menlo Park, California; London; Don Mills, Ontario.

Ohio Agricultural Research & Development Center. 1968. 1967 Ohio Farm Income. Ohio Agricultural Research and Development Center, Wooster, Ohio.

Spiegel, Murray R. 1961. Theory and Problems of Statistics. McGraw-Hill Book Company, New York, St. Louis, San Francisco, Toronto, Sydney.

U.S.D.A. 1968. Livestock and Meat Situation. Nov. Issue. Economic Research Service. Washington, D. C.

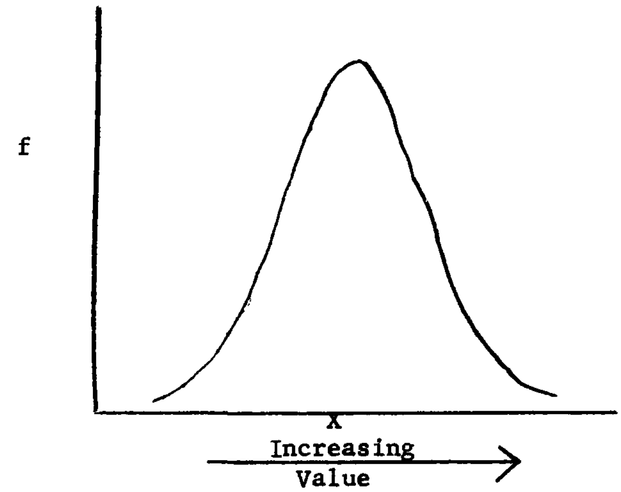
U.S.D.A. 1968. Livestock snd Meat Situation. Nov. Issue. Economic Research Service, Washington, D. C.

Walker, Helen M. and Joseph Lev. 1943. Elementary Statistical Methods. Holt, Rinehart and Winston, New York.

DATA DISTRIBUTION

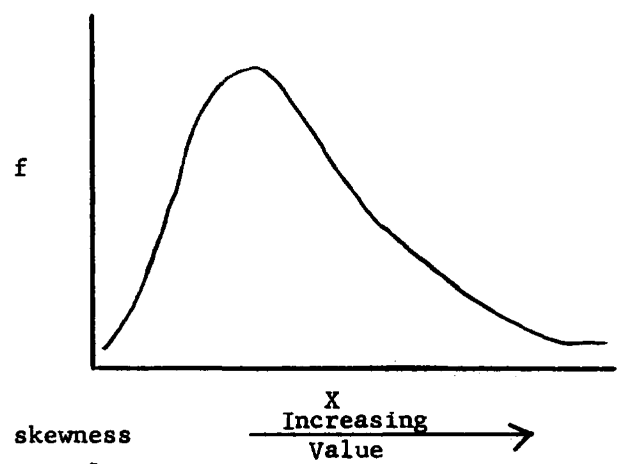
Data if arranged in a frequency distribution and graphed will yield several types of curves and they may be subdivided as follows:

1. Symmetrical Normal Distribution

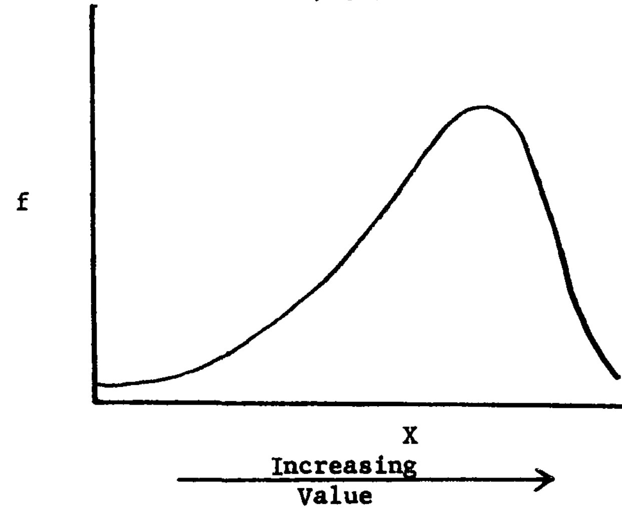


2. Skew

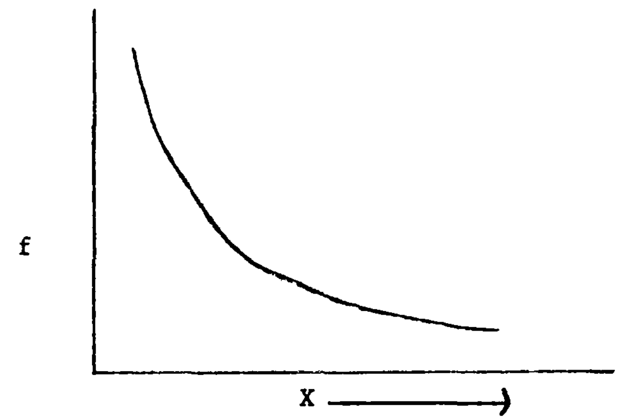
a) Normal distribution-positive skewness



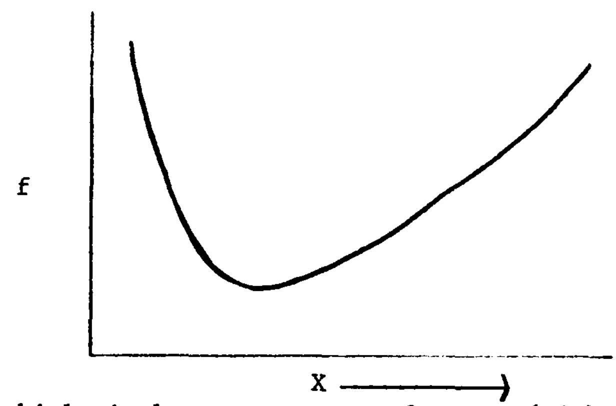
b) Normal distribution-negative skewness



c) J-curve-positive skewness (can also be negative)



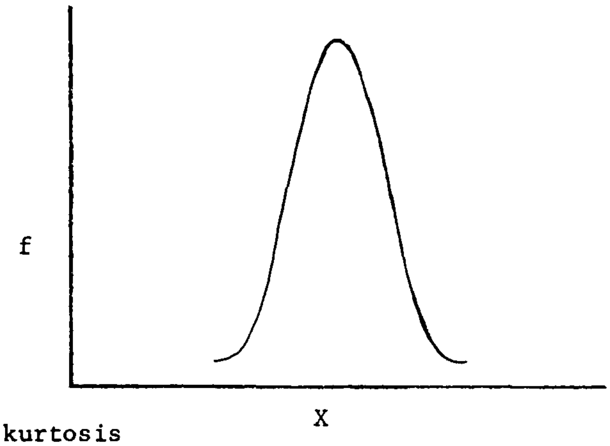
d) U-shape-negative skewness (can also be positive)



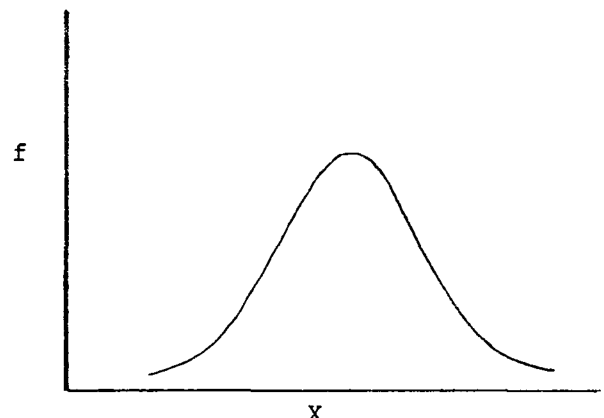
Fortunately, in most biological systems a normal curve (with slight skewness in either direction) is usually approximated. This fact is quite important in simplifying the statistic in the biological area.

The normal (symmetrical, bell-shaped, law or error) curve can still vary and retain its symmetrical properties as shown in the following examples:

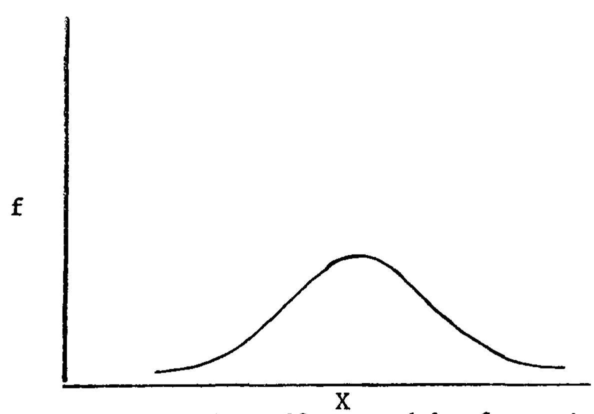
1. Positive kurtosis



2. Zero kurtosis



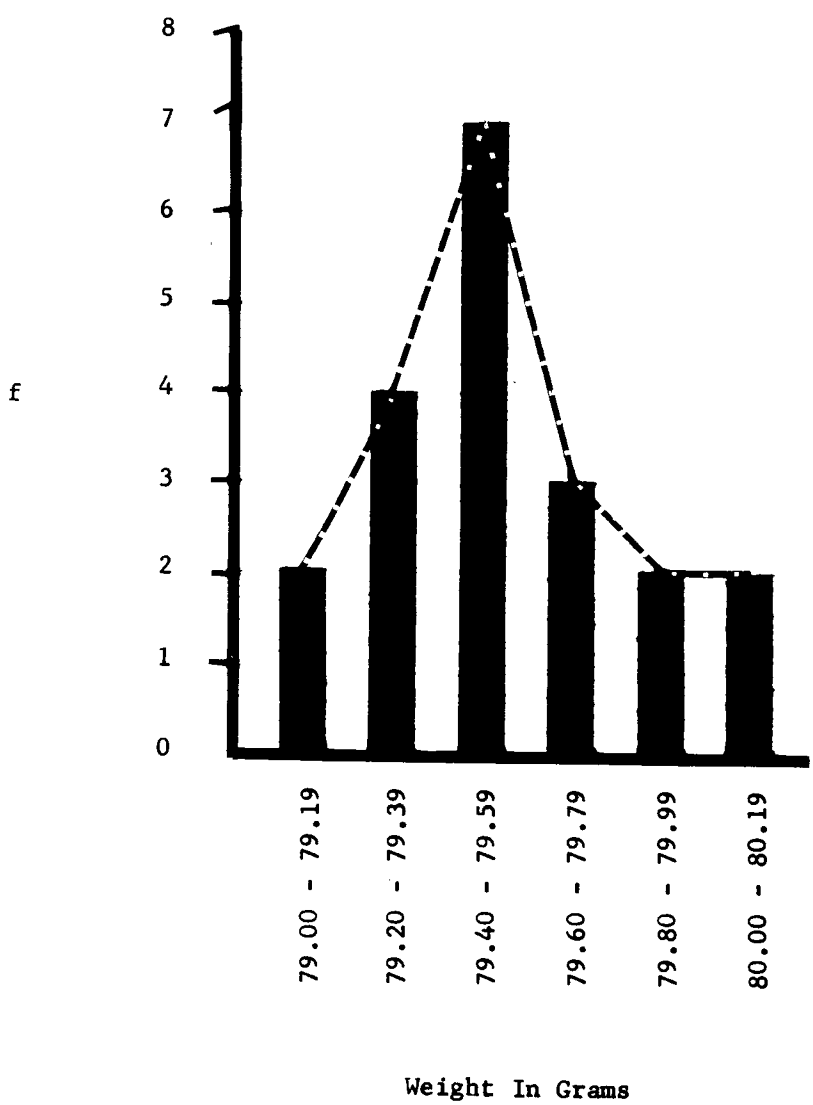
3. Negative kurtosis



To check data distribution, 20 ground beef patties were weighed as they came from a patty machine and their weight was classified as follows:

Weight in grams	Frequency	f
79.00-79.19	11	2
79.20-79.39	1111	4
79.40-79.59	~~1111~~ 11	7
79.60-79.79	111	3
79.80-79.99	11	2
80.00-80.19	11	2
		20

This distribution was then plotted on a graph to yield the following:



The perfect normal curve was not obtained (it very seldom is) but the data should come closer to approximating it as the sample size is increased. (For example, to 500 and to 1000 samples.)

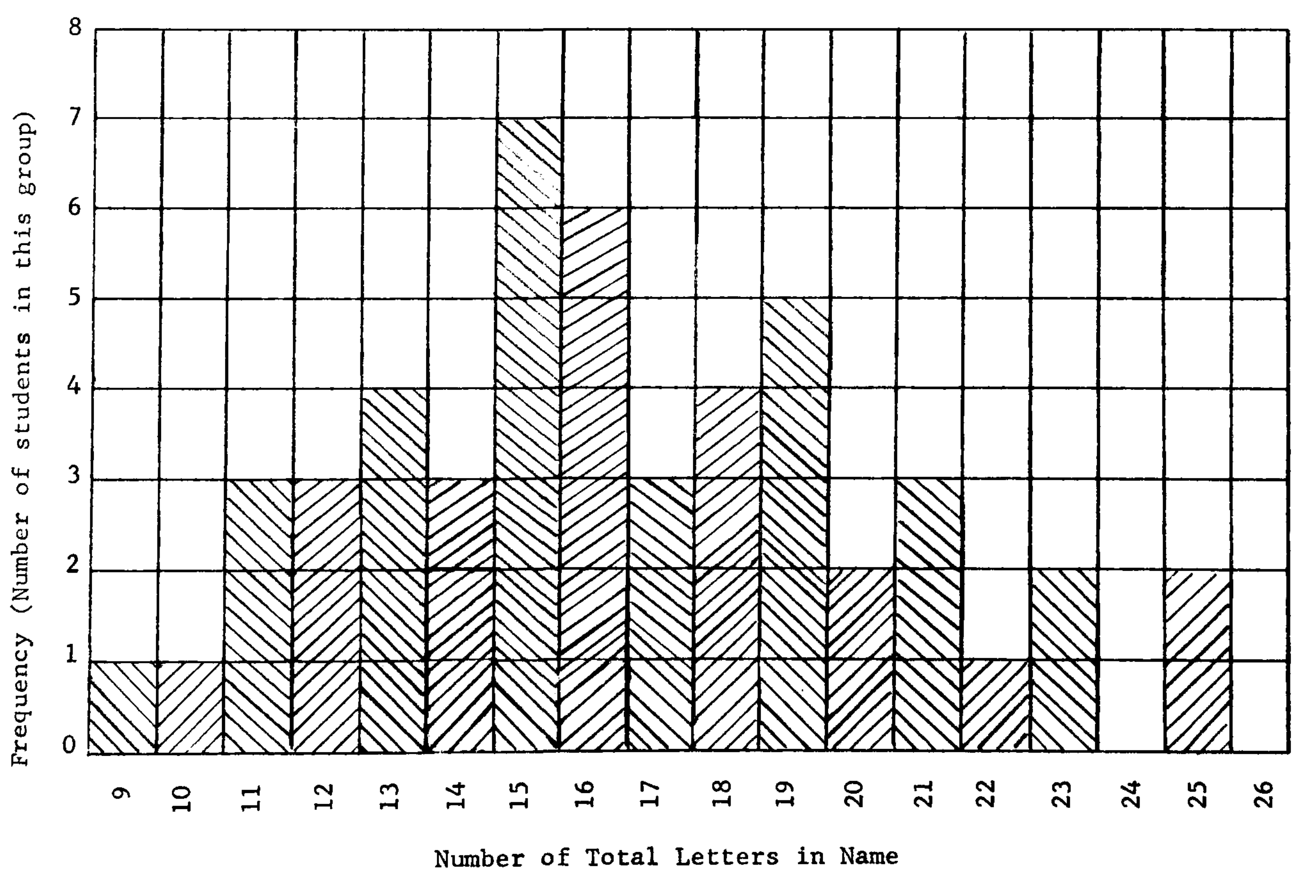
If the machine sometimes does not fill the mold completely a negative skewness would result and if the machine often fills the mold 3/4 full a positive skewness would be expected.

Most biological data will follow an approximate normal distribution curve but skewness is often expected in the following type of examples.

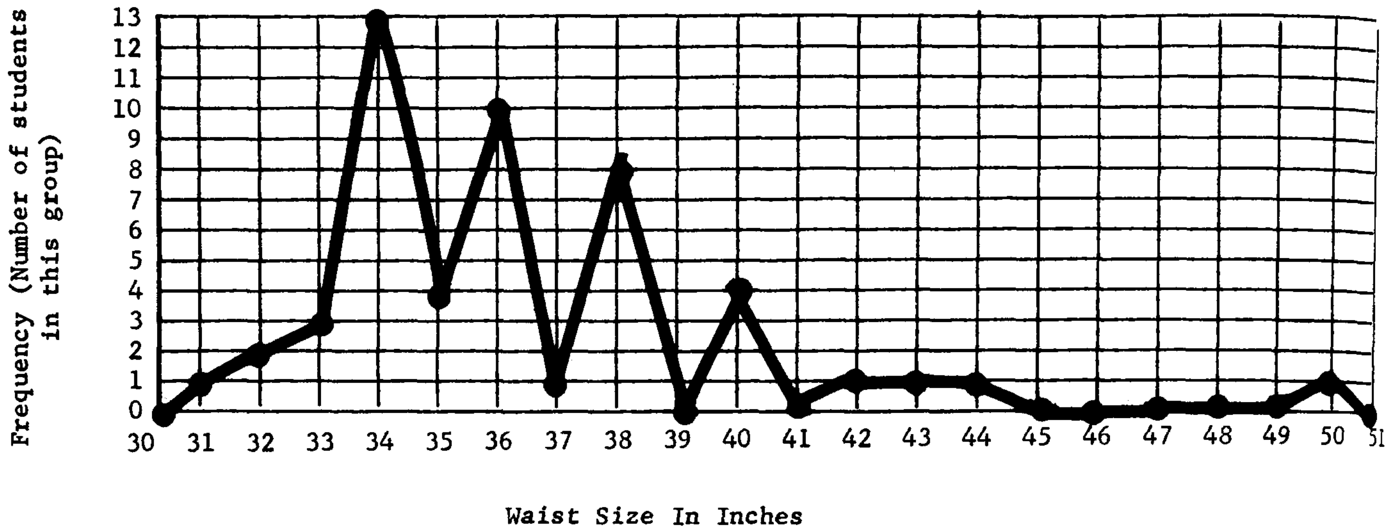
		Skewness	Horizontal Scale
1.	Easy test	- Positive	Number of wrong answers
2.	Average test	- Symmetrical	Number of wrong answers
3.	Difficult test	- Negative	Number of wrong answers
4.	Number of children per family	- Positive	0 to 20 children
5.	Age at marriage	- Positive	0 to 120 years

Examples of Distributions Collected From Class Members

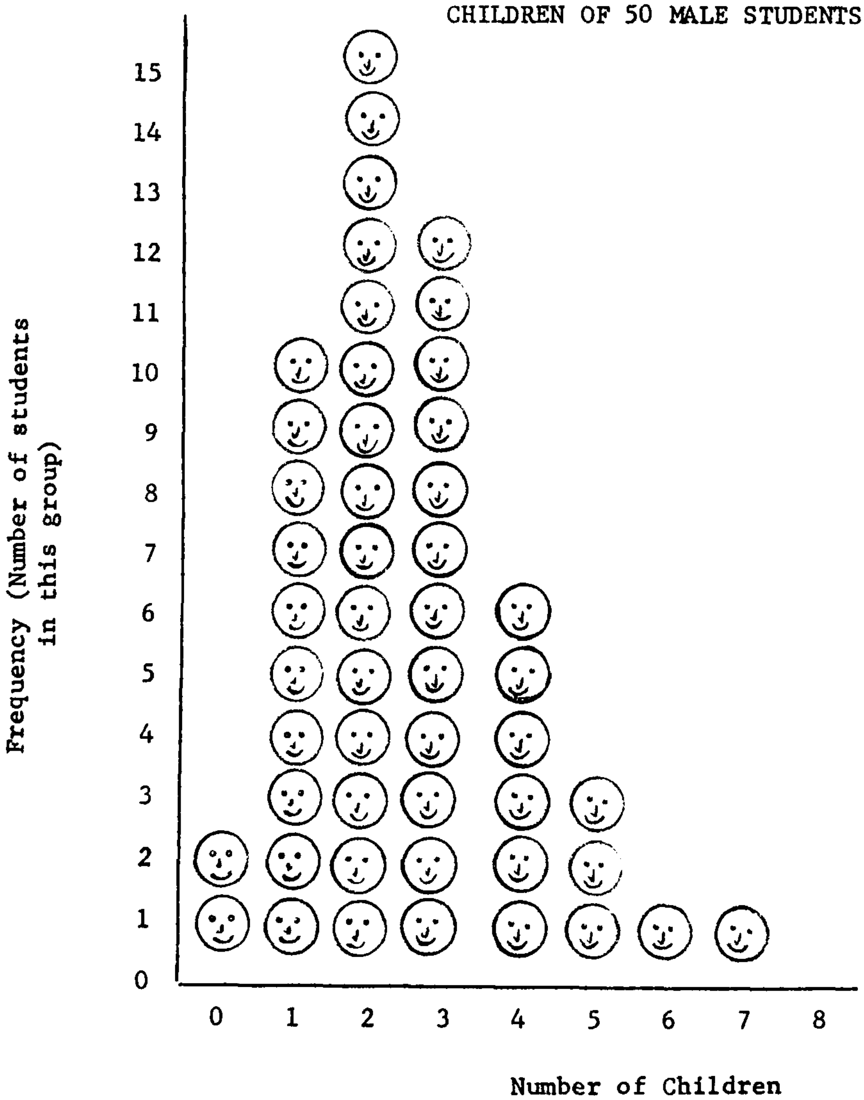
FREQUENCY DISTRIBUTION OF NUMBER OF LETTERS IN NAMES OF 50 PEOPLE



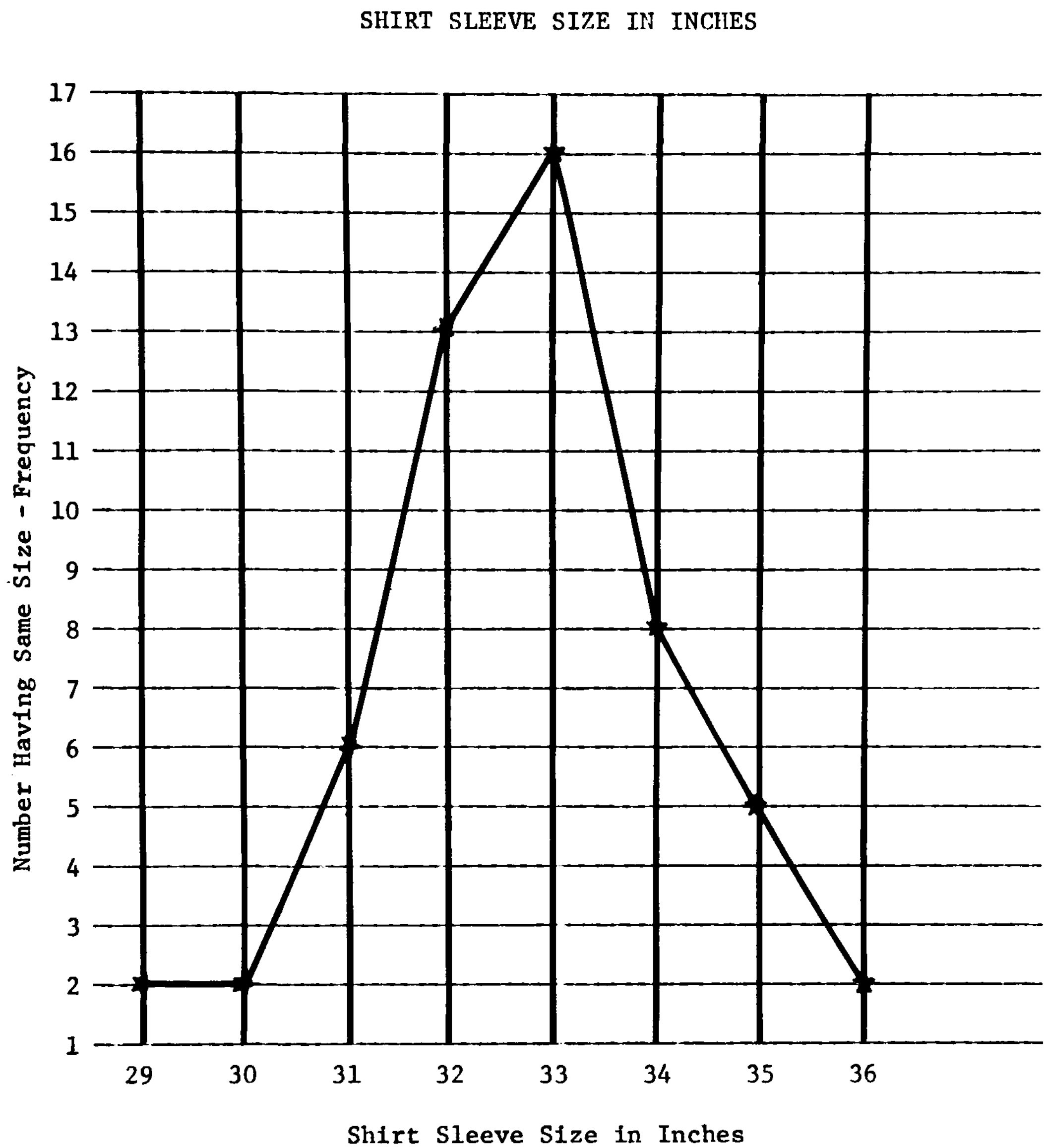
FREQUENCY DISTRIBUTION OF WAIST SIZE IN INCHES



FREQUENCY DISTRIBUTION OF NUMBER OF CHILDREN OF 50 MALE STUDENTS

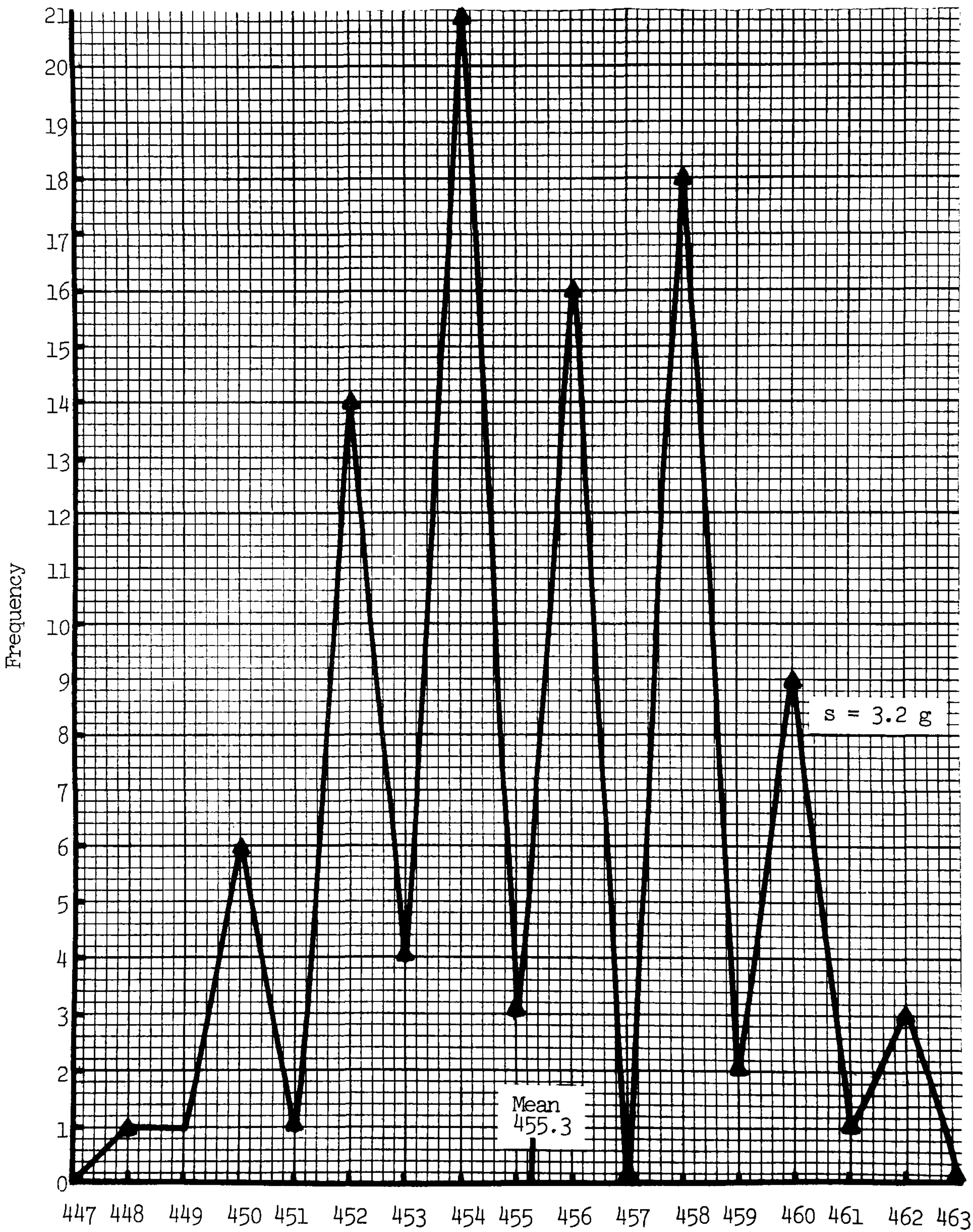


How would you classify the distributions collected from the class? What caused the skewness where it was observed?



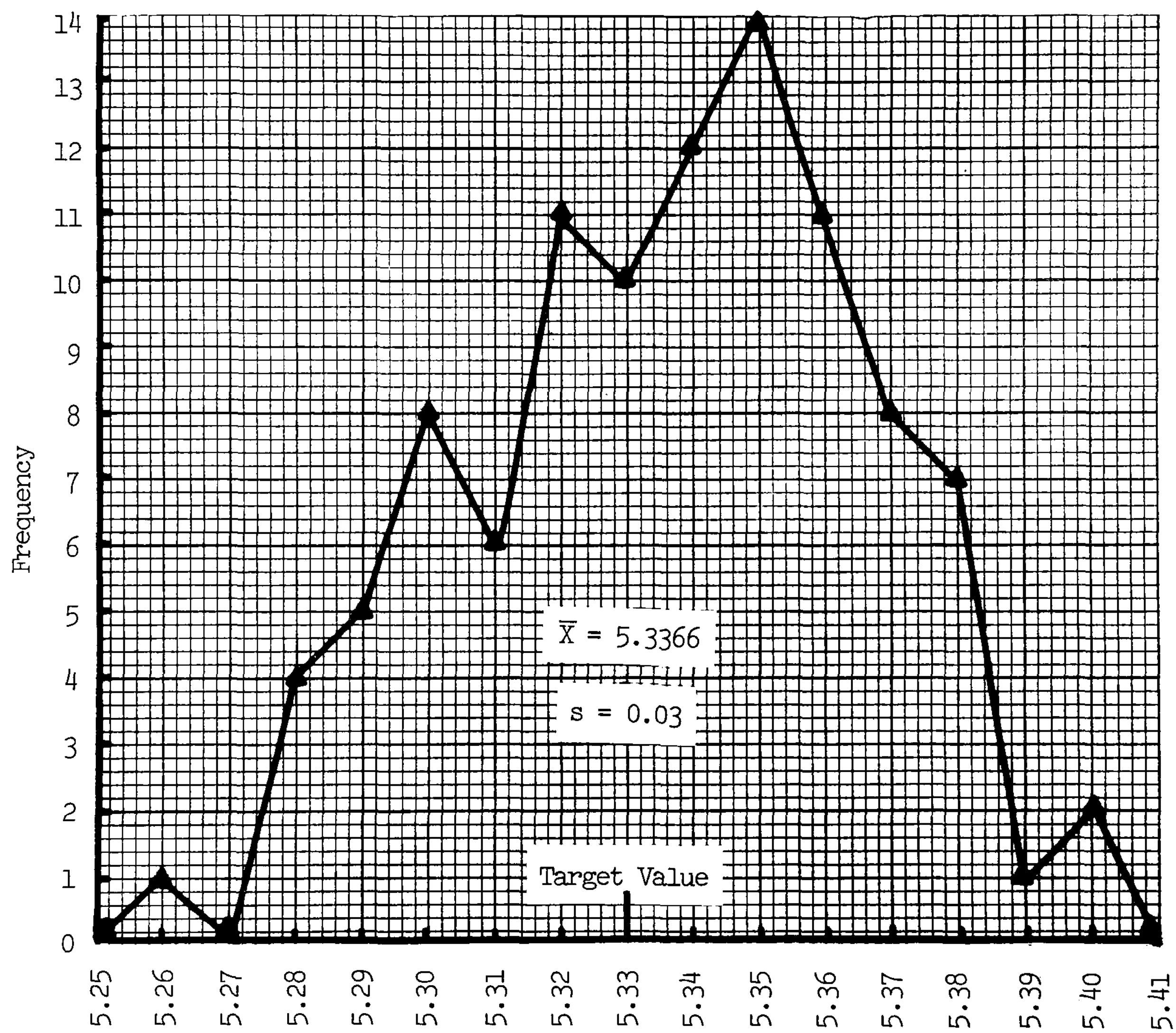
The following graphs show the types of distribution often encountered in the meat area. Notice that it also is close to normal.

Frequency distribution of 6/1 beef patties



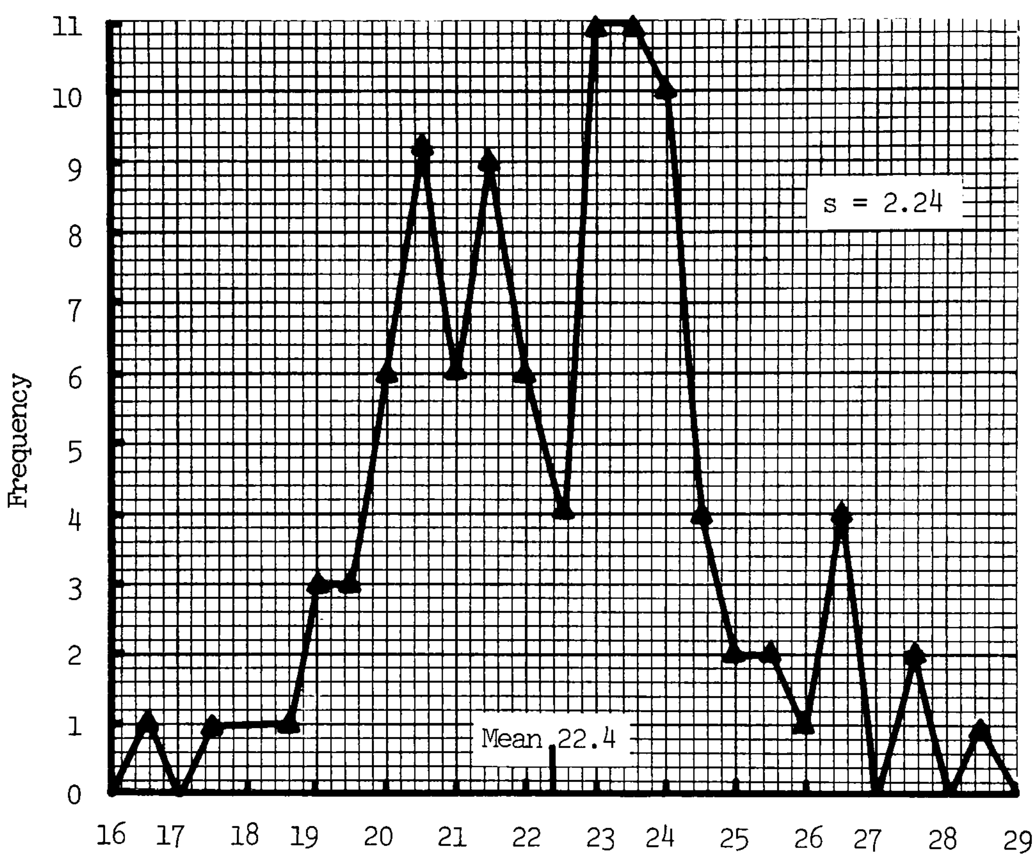
Weight in grams of 6/1 beef patties

Frequency distribution of 3/1 round steaks



Oz. of 3/1 round steak

Frequency distribution of fat in ground beef



% fat in regular ground beef

References

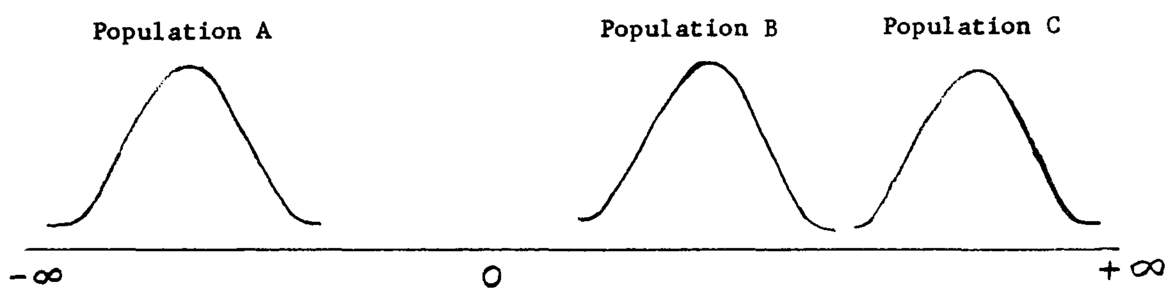
Bartz, Albert E. 1963. Elementary Statistical Methods for Educational Measurement. Burgess Publishing Company, Minneapolis, Minnesota.

Smith, G. Milton. 1962. A Simplified Guide to Statistics. Holt, Rinehart and Winston, Inc., New York.

Treloar, Alan E. 1939. Elements of Statistical Reasoning. John Wiley & Sons, Inc., New York - London, Chapman & Hall, Limited.

MEASURE OF CENTRAL TENDENCY

To describe or define a population the initial task is to calculate or measure where the center of this population lies. Many populations may have centers anywhere from $-\infty$ (minus infinity) through 0, to plus ∞ as illustrated by populations A, B, & C.



Some populations may have only positive centers and would be represented by population B and C. The centers of population A, B and C are different and information about the location of these centers would be valuable in defining these populations (also necessary when comparing 2 different populations).

Measures of central tendency will supply information on the location of the center of a population. There are several types of measures and a few of the more common are listed below:

1. Arithmetic mean or mean
2. Median
3. Mode
4. Geometric mean
5. Harmonic mean

Each of these measures of central tendency has its advantages and disadvantages depending on the data and the intended use of the results.

ARITHMETIC MEAN

The arithmetic mean or mean is the most common measure of central tendency. $\overline{X}$ (X bar) is the symbol used to designate the mean. It is calculated by adding the variable values and dividing by the number of variables. For example, the following ham weights are obtained.

Ham Number	Ham Weight X	Statistical Designation X_i
1	12	X_1
2	16	X_2
3	14	X_3
4	12	X_4
5	14	X_5
Total	68	

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{n} = \frac{12 + 16 + 14 + 12 + 14}{5} = \frac{68}{5} = 13.6$$

This can also be written

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$\sum$ = sum of

$\sum_{i=1}^{n} X_i$ = sum of $X_1 + X_2 + X_3 \ldots X_n$

If there were a large number of hams and they were listed in a frequency table the calculation could be simplified by the following method:

Class Mid Point	No. of hams of this weight f	Total weight of hams in this group
12	$f_1 = 2$	24
13	$f_2 = 0$	0
14	$f_3 = 2$	28
15	$f_4 = 0$	0
16	$f_5 = 1$	16
Total	5	68

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + f_3X_3 + f_4X_4 + f_5X_5}{f_1 + f_2 + f_3 + f_4 + f_5}$$

$$\bar{X} = \frac{2(12)+0(13)+2(14)+0(15)+1(16)}{5} = \frac{24+28+16}{5} = \frac{68}{5} = 13.6$$

may also be written

$$\bar{X} = \frac{\sum_{i=1}^{n} f_iX_i}{\sum_{i=1}^{n} f_i}$$

The arithmetic mean has several properties that are quite helpful in subsequent statistical manipulation. These properties are as follows:

1. If each original value is compared to the mean (deviation) the sum of these comparisons will equal zero. For example:

Ham Number	Ham Weight X	Ham Mean $\bar{X}$	Deviation $x = (X-\bar{X})$
1	12	13.6	-1.6
2	16	13.6	+2.4
3	14	13.6	+0.4
4	12	13.6	-1.6
5	14	13.6	+0.4
	$\sum X =$ 68		$\sum x =$ 0

This calculation is often used to check the correctness of the calculated mean.

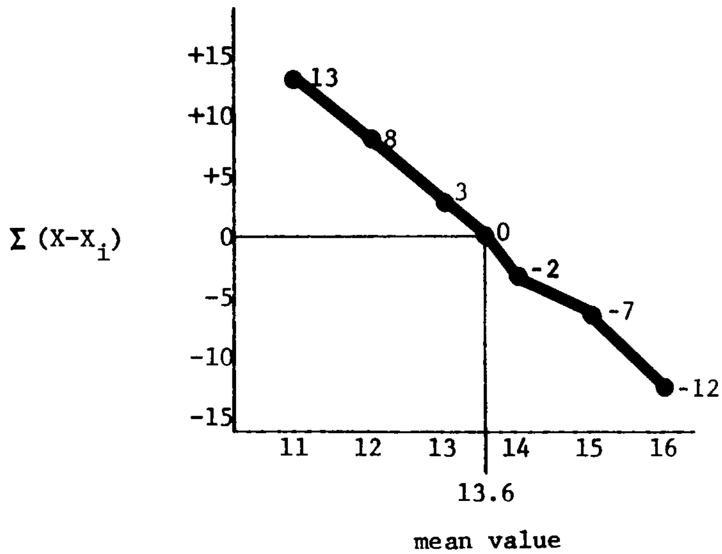
2. If the deviations from the mean are squared and totaled this will be the smallest number that may be obtained using this procedure. For example, if we take the values 13.0, 13.6, and 14.0 and find the deviations from these values, square these deviations and then total the results, the smallest total value will be obtained with 13.6 (mean).

Ham #	Ham Wt. X	Mean $\bar{X}$	Deviation $x=(X-\bar{X})$	Deviation squared x^2	Y	Deviation $y=(X-Y)$	Deviation squared y^2	Z	Deviation $z=(X-Z)$	Deviation squared z^2
1	12	13.6	-1.6	+2.56	13.0	-1.0	+1.00	14.0	-2.0	+4.00
2	16	13.6	+2.4	+5.76	13.0	+3.0	+9.00	14.0	+2.0	+4.00
3	14	13.6	+ .4	+ .16	13.0	+1.0	+1.00	14.0	0	0
4	12	13.6	-1.6	+2.56	13.0	-1.0	+1.00	14.0	-2.0	+4.00
5	14	13.6	+ .4	+ .16	13.0	+1.0	+1.00	14.0	0	0
	$\Sigma X=68$		$\Sigma x=0$	$\Sigma x^2=11.20$		$\Sigma y=+3.0$	$\Sigma y^2=13.00$		$\Sigma z=-2.0$	$\Sigma z^2=12.00$

If values other than the mean (13.6) are used, for example -- 13.0 or 14.0, the sum of the deviation (Σy and Σz) will not equal zero (+3 and -2 respectively) indicating that these values are not the true mean. If the deviations are squared and summed (Σx^2 - sum of squares) the smallest result will be obtained when the mean is used. For example:

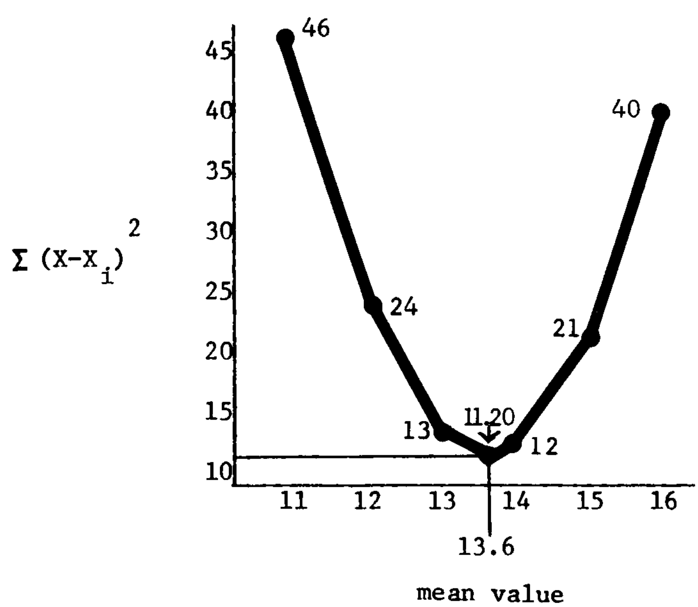
Values Used	Calculations	Sum of Deviations	Sum of Deviations Squared
11	not shown	13	45.00
12	not shown	8	24.00
13	in previous table	3	13.00
13.6 (mean)	in previous table	0 (zero)	11.20 (smallest)
14	in previous table	-2	12.00
15	not shown	-7	21.00
16	not shown	-12	40.00

If the sum of deviations were plotted against values used, the following graph would result:



Values below the mean would yield a positive deviation, the mean will yield a zero deviation and values above the mean will yield a negative deviation.

If the sum of deviations squared were plotted against values used, the following graph would result:



Values below and above the mean will yield larger sum of deviations squared than will the mean value. The mean values will yield the smaller sum of deviations squared possible.

The reduction in the sum of the deviations from a statistic is often used as a measure of how well the statistic describes the sample.

If the calculation of the deviation from the mean $\left[(X - \bar{X}) \text{ or } x\right]$ or the $\sum$ of deviation squared is desired and the data are expressed in a frequency table the following procedure may be used for this calculation:

Ham Weight Classes X_i	No. of hams of this wt. f	Ham mean $\bar{X}$	Class deviation from mean $X_i-\bar{X}$	Ham deviation from mean $f(X_i-\bar{X})$	Ham deviation from mean squared $f(X_i-\bar{X})^2 =$
12	2	13.6	-1.6	-3.2	2(2.56) = 5.12
13	0	13.6	-0.6	0	0(0.36) = 0
14	2	13.6	+0.4	+ .8	2(0.16) = 0.32
15	0	13.6	+1.4	0	0(1.96) = 0
16	1	13.6	+2.4	+2.4	1(5.76) = 5.76
				$\sum = 0$	$\sum = 11.20$

The values agree with the same data calculated from the raw data rather than from a frequency distribution table.

3. The mean is very sensitive to extreme values and this may be shown by the following example:

	Live Hog Weights X_1		Live Hog Weights X_2
	200 pounds		200 pounds
	190		190
	210		210
	220		220
	180		880
$\sum X_1 =$	1000	$\sum X_2 =$	1700
$\bar{X}_1 =$	200	$\bar{X}_2 =$	340

The change in the weight of one hog to an extreme value almost doubled the average weight.

4. If repeated samples are taken from a population the mean in contrast with other measures of central tendency will have the most stable value (less fluctuation between samples).

MEDIAN

The median is another measure of central tendency and if values are ranked from low to high (array) the median is the middle value. It is calculated in one of the following ways.

1. Calculations of the median if the number of observations are odd: After ranking the observation in an array the median is the middle value.

   Example - weights of pork carcasses; 102, 105, 112, 107, 120

   Arrange in an array; 102, 105, 107, 112, 120

   The middle weight is 107 with two weights above it and two weights below it.

2. Calculations of the median if the number of observations are even: If the number of observations are even the median is the arithmetic average of the two middle values.

   Example - weights of pork carcasses; 110, 108, 125, 115, 106, 118

   Arrange in an array; 106, 108, 110, 115, 118, 125

   The two middle values are 110 and 115

   Their mean is = (110 + 115)/2 = 112.5

   The median is 112.5 and there are 3 values above the median and 3 values below it.

3. Calculation of the median when several scores have the same value as the middle score:

   The ham data used in the mean example would fit into this category.

   Example of these ham weights in an array; 12, 12, 14, 14, 16

   If these data are considered as a frequency table with class intervals of one and class limits at .5 (example - 12 would be 11.5

to 12.4) the following formula may be used to calculate the median. In this example by visual observation the median is between 13.5 and 14.4.

$$\text{Median} = \begin{array}{l}\text{lower class limits}\\ \text{of class contain-}\\ \text{ing median}\end{array} + \left(\frac{\dfrac{\text{No. of observations in total sample}}{2} - \begin{array}{l}\text{Sum of all fre-}\\ \text{quency below}\\ \text{median class}\end{array}}{\text{frequency of median class}} \right) \begin{array}{l}\text{Class}\\ \text{interval}\end{array}$$

$$= 13.5 + \left(\frac{\frac{5}{2} - 2}{2} \right) \cdot 1 = 13.5 + \left(\frac{2.5-2}{2} \right) \cdot 1 = 13.5 + \left(\frac{.5}{2} \right) \cdot 1$$

$$= 13.5 + .25 = 13.75$$

This value is different from the mean of 13.6.

Properties of the Median

1. The median will divide a histogram into two parts each having an equal area. The number of scores below the median equals the number above it.
2. The median is more representative of the sample when the sample contains a few non-typical large or small observations.
3. Each observation will have the same influence on the median regardless of its size. (Not the case with the mean).

<u>MODE</u>

The mode of a population is the observation that occurs the most frequently or that is the most common. The mode class interval is the highest point on a histogram. A sample may have the following:

1. One mode - the number of observations falling in one class interval exceeds the number in any other class interval. Example, ham weights.

Ham Weight Classes		f	
12	-	3	
13	-	6	
14	-	12	- mode is 14 (not 12)
15	-	9	
16	-	4	

2. Two modes - Two classes have the same number of observations which are greater than any other class frequency.

Ham Weight Classes		f		
12	-	3		
13	-	10	- mode	$\frac{13+15}{2} =$
14	-	7	average	14 mode
15	-	10	- mode	
16	-	2		

The average of these class intervals is considered the mode.

3. No mode - No class interval has a majority of observations.

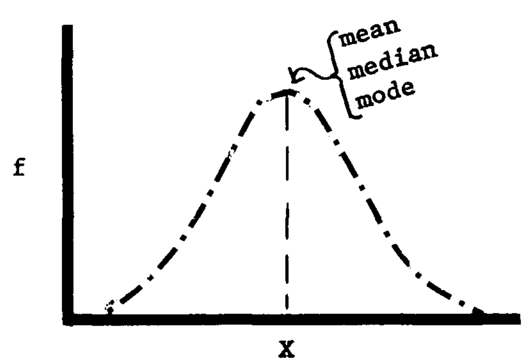
Ham Weight Classes		f	
12	-	1	
13	-	0	No mode
14	-	1	
15	-	1	
16	-	1	

There is a mathematical calculation to determine the exact point of a mode within a frequency class but it is used primarily in theoretical work.

The primary advantage of a mode is the ease of obtaining this statistic. As a measure of central tendency the mode is a rough approximation and will sometimes suffice.

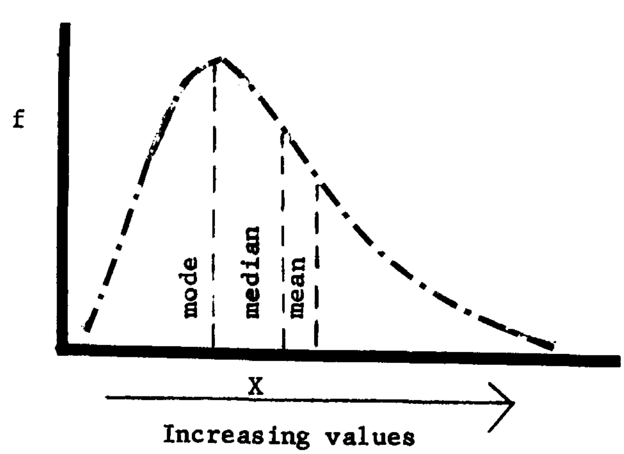
<u>Location of the mean, median and mode</u>

1. In a symmetrical curve:



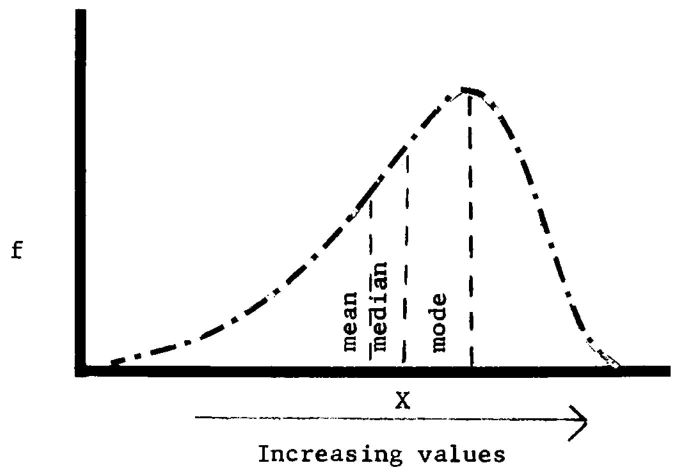
The mean, median, and mode all have the same value.

2. In a positive skewed distribution:



In a positively skewed distribution the mode has the smallest value, the median has a larger value, and the mean has the largest value. The distance between the mode and median is about twice as great as between the median and the mean.

3. In a negative skewed distribution:



In a negatively skewed distribution the order of measures of central tendency is reversed. The mean has the lowest value, the median a higher value, and the mode the highest value. Again, the distance between the mean and median is about one-half the distance between the median and the mode.

Other measures of central tendency have specialized uses and will be mentioned only briefly.

GEOMETRIC MEAN

This mean is determined by multiplying the observations together and then taking the root (equal to the number of observations) of this product. An example of the geometric mean of 1, 1, 1, 5, 25, and 125 is as follows:

$$\text{Geometric mean} = \sqrt[n]{(X_1)\quad(X_2)\ \dots\ (X_n)}$$

$$\text{Geometric mean} = \sqrt[6]{1 \times 1 \times 1 \times 5 \times 25 \times 125}$$

$$= \sqrt[6]{15625} = 5$$

This type of mean is often used when the variable is changing at a constant rate with respect to time. The average growth of microorganisms on a meat product would be an example of such an application.

HARMONIC MEAN

The harmonic mean is computed as follows:

1. The reciprocal of each observation is taken
2. The mean is computed on #1
3. The reciprocal of this mean (#2) is taken

$$\text{Harmonic mean} = \frac{1}{\frac{\Sigma(1/X)}{n}} = \frac{n}{\Sigma(1/X)}$$

The harmonic mean of 2, 3, 6

$$\text{Harmonic mean} = \frac{3}{\frac{1}{2} + \frac{1}{3} + \frac{1}{6}} = \frac{3}{\frac{3}{6} + \frac{2}{6} + \frac{1}{6}} = \frac{3}{\frac{6}{6}}$$

$$= \frac{3}{1} = 3$$

This type of mean is often used when problems involving time, rate and distance (or work) are calculated. The harmonic mean is often used to calculate the following:

1. a) Data are given in <u>minutes to complete</u> a job for each individual worker. Example:

Worker	Minutes to bone a ham
Number 1	5 minutes per ham
Number 2	6 minutes per ham

b) The average number of minutes per worker to complete a job for the plant is required. Example:

$$\text{Harmonic mean} = \frac{n}{\Sigma(1/X)} = \frac{2}{\frac{1}{5} + \frac{1}{6}} = \frac{2}{\frac{11}{30}}$$

$$= \frac{2}{1} \cdot \frac{30}{11} = \frac{60}{11} = 5.45 \text{ minutes per ham (Not 5.5)}$$

2. a) Data are in number of pounds of meat that can be purchased for sausage for a given sum of money. Example:

Ingredient	Cost	Quantity bought
Ingredient #1	2 pounds per $1.00	$1.00
Ingredient #2	3 pounds per $1.00	$1.00

b) Wanted average price of meat delivered. (Weighted average would also work here.)

$$\text{Harmonic mean} = \frac{n}{\Sigma(1/X)} = \frac{2}{\frac{1}{50} + \frac{1}{33\ \&\ 1/3}} = \frac{2}{2/100 + 3/100}$$

$$= \frac{2}{1} \cdot \frac{100}{5} = \frac{200}{5} = 40\text{¢ (Not 41.7¢)}$$

<u>Comparison of means</u>

Comparison of the means of a set of positive numbers

1. Arithmetic mean - Is greater than or equal to the geometric mean and the harmonic mean.

2. The geometric mean is less than or equal to the arithmetic mean and greater than or equal to the harmonic mean.

3. The harmonic mean is less than or equal to the geometric mean and the arithmetic mean.

Harmonic mean (Smaller)	$\leq$	Geometric mean	$\leq$	Arithmetic mean (Larger)

Sample Problems

1. Mean (arithmetic mean) - a measure of central tendency.

Example:

sample observations X
2
3
4
4
2
$\sum X = 15$

The mean ($\bar{X}$) is obtained by adding the individual observations and dividing by the number of these observations.

$$\bar{X} = \frac{\sum X}{n}$$

$$\bar{X} = \frac{15}{5} = 3$$

The mean can be checked as follows:

X	$\bar{X}$	$x = (X - \bar{X})$
2	3	-1
3	3	0
4	3	+1
4	3	+1
2	3	-1
$\sum X = 15$		$\sum (X - \bar{X}) = 0$

If the mean ($\bar{X}$) is substracted from the observation (X) a term called deviation from the mean $[(X - \bar{X}) = x]$ is obtained. If these deviations are summed and if this sum is zero the mean is correct.

Find the mean of the following group of samples. Check this mean by totaling the deviations to determine if it is correct before you check the answers listed below.

A	B	C	D	E	F	G	H	I	J
4	7	6	3	2	0	2	3	10	2
6	8	3	3	2	0	3	3	100	4
	9	3	3	3	4		3		6
	8		3	3	4		3		8
			8	0					

List 4 properties of the mean.

Answers. (A-5, B-8, C-4, D-4, E-2, F-2, G-2.5, H-3, I-55, J-5.)

2. Median - a measure of central tendency. To obtain the median the observations are placed in an array and the center observation is selected.

<u>Example</u>:

Odd number of observations

observation X	in an array	
6	1	3 below
7	2	
3	3	
2	6	← center value in the median
1	7	3 above
9	9	
11	11	

Even number of observations

observation X	in an array	
1	1	2 below
9	2	
3	3	$\frac{3+6}{2}$ = 4.5 is the median
7	6	
6	7	2 above
2	9	

Find the median of the following groups of observations.

A	B	C	D	E	F	G	H	I	J
1	2	9	2	9	1	9	2	2	2
9	6	7	2	7	2	8	2	2	0
7	4	3	6	9	3	7	1	2	0
	1	2	6	3	4	6	1		3
		8	3	4	5	5	3		4
		1					3		4
							4		5
							6		
							9		
							8		
							6		
							9		
							8		

List 3 properties of the median

Answers (A-7, B-3, C-5, D-3, E-7, F-3, G-7, H-4, I-2, J-3)

3. Mode - a measure of central tendency. Construct a frequency table. The mode is the observed value that occurs the most frequent. The mode is the high point on a frequency graph.

Example:

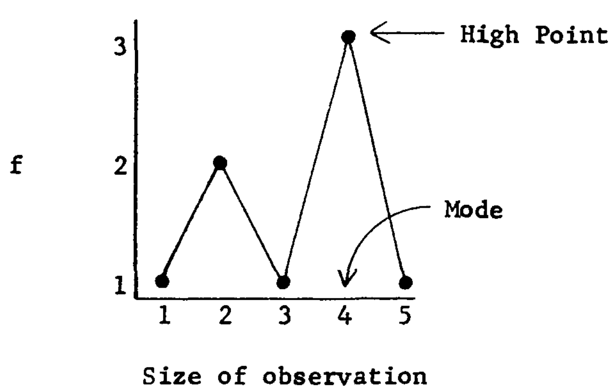
observations X
2
3
4
4
2
1
5
4

$\Big\}$ n=8 ← equal → $\sum f=8$

Plot in a frequency table

size of observations: Mid Point of the class	frequency of occurrence	
1	1	
2	2	
3	1	
4	3	Highest frequency, mode value is the mid point of this class - 4
5	1	
	$\sum f=8$	

Plotted



Find the mode of the following groups of samples.

A	B	C	D	E	F	G	H	I	J
6	1	2	1	0	1	1	1	100	2
7	3	8	9	1	2	3	2	125	6
3	4	7	3	2	2	4	2	123	7
6	2	7	8	4		2	3	125	6
2	1	6	4	0			2	127	3
1	8		2	3			2		1
	9		6				3		2
	1		3				3		7
	2		1						7

Name an advantage and a disadvantage of the mode.

Answers. (A-6, B-1, C-7, D-2, E-0, F-2, G-No Mode, H-2, I-125, J-7)

References

Chou, Ya-Lun. 1963. Applied Business and Economic Statistics. Holt, Rinehart and Winston, New York, Chicago, San Francisco, Toronto, London.

Haber, Audrey and Richard P. Runyon. 1969. General Statistics. Addison-Wesley Publishing Company, Reading, Massachusetts; Menlo Park, California; London; Don Mills, Ontario.

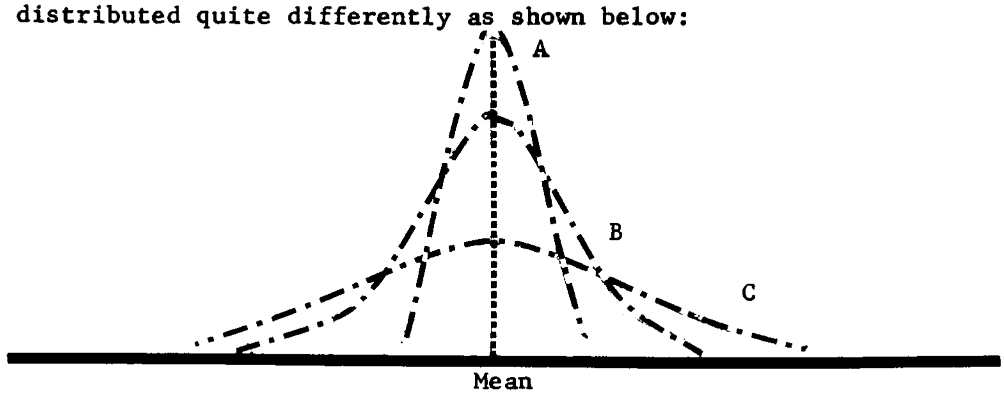
Smith, G. Milton. 1958. A Simplified Guide to Statistics. Holt, Rinehart and Winston, Inc., New York.

Spiegel, Murray R. 1961. Theory and Problems of Statistics. McGraw-Hill Book Company, New York, St. Louis, San Francisco, Toronto, Sydney.

Walker, Helen M. and Joseph Lev. 1958. Elementary Statistical Methods. Holt, Rinehart and Winston, New York.

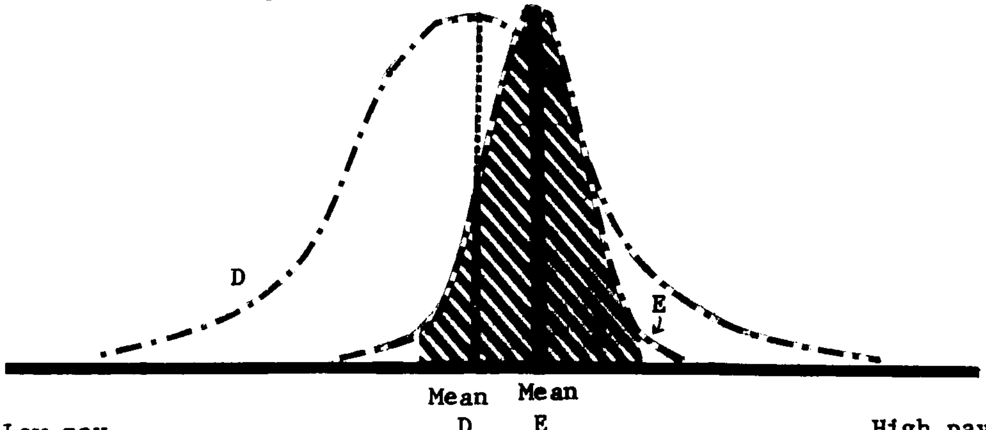
MEASURES of VARIABILITY

After a measure of central tendency has been determined, one more statistic is needed to adequately describe the population and this is a measure of the amount of the variability. Populations with the same mean may be distributed quite differently as shown below:



These populations may represent the final weights of a pound of weiners produced by 3 packers--A, B and C. If the mean is one pound, all packers are producing a product that, on the average, is the correct weight. Packer A is producing a much more uniform product than B or C and most of his (A) packages would weigh very near the mean weight (one pound). On the other extreme, packer C is producing a product that also averages one pound but his product is not uniform and his product weight will vary both above and below the pound mark frequently. Certainly packer A is doing a much better job than packer C even though both of their products will average one pound.

When comparing populations with different means an indication of the variability is needed to properly interpret the difference between the populations. For example:



Company E pays its employees on the average more than Company D. Company D has more variability in its pay scale than Company E. If an employee only knew the average pay scale he would prefer to work for Company E. A good portion of Company D employees make more money than any of the Company E employees; therefore, if the worker had a high skill he probably would receive more pay with Company D. If on the other hand he was low skilled, he would be better off with Company E.

It takes 2 values to adequately describe a population.

1. A measure of central tendency
2. A measure of variability

Five measures of variability are commonly used in statistics and they are:

1. Range
2. Mean deviation
3. Variance
4. Standard deviation
5. Coefficient of variation

RANGE

The range is normally calculated by subtracting the smallest value from the largest value. Sometimes both the smallest and largest values are reported. For example, the following are live hog weights as recorded by a stockyard: 210, 180, 220, 240, 195, 850 pounds

Range or R is: 850 - 180 = 670 or 180 to 850 pounds

Properties of the range

1. The main advantage of the range is that it is easy to obtain.
2. The range is not as precise as some of the following measures of variability because it is based on only the 2 extreme observations out of the sample and the other observations have no influence on it. An extreme value (like 850 pounds in the

previous example) in either direction will give a false impression of the total degree of variability.

3. The more samples that are drawn from a population the greater the possibility of obtaining an extreme sample and increasing the size of the range. If samples are drawn from the same population, the larger the sample size the larger the expected range.

MEAN DEVIATION

The mean deviation as a measure of variability is calculated by adding the deviations from the mean, without regard to sign, and dividing by the number of observations.

The following ham data was used in calculating the mean:

Ham Number	Ham Weight X	Ham Mean $\bar{X} = \Sigma X/n$	Deviation From the mean $x = X-\bar{X}$
1	12	13.6	-1.6
2	16	13.6	+2.4
3	14	13.6	+0.4
4	12	13.6	-1.6
5	14	13.6	+0.4
n=5	$\Sigma X = 68$		$\Sigma x = 0$

Sum, without regard to sign = 6.4

$$\text{Mean Deviation} = \frac{\Sigma|X - \bar{X}|}{n}$$

2 straight lines indicate an absolute value or a value without regard to sign

$$\text{Mean Deviation} = \frac{6.4}{5} = 1.28$$

This value measures the average distance each observation lies from the mean value. The mean (13.6 is this example) plus or minus the mean deviation (1.28 in this example) should give a pair of values that includes approximately 50% of the original observations.

13.6 (mean) + 1.28 (mean deviation) = 14.88

13.6 (mean) - 1.28 (mean deviation) = 12.32

The original observations that are underlined are within the values of 12.32 to 14.88.

Ham No.	X Ham Weight
1	12
2	16
3	<u>14</u>
4	12
5	<u>14</u>

2 out of 5 or 40% of the original observations are within these values instead of the expected 50%.

As the mean deviation increases in magnitude there is more variability in the observations. These <u>principles</u> are used in most measures of variability even though the mean deviation has found only limited use.

<u>Properties of mean deviation</u>:

The mean deviation has 2 serious faults:

1. It gives poor estimates as to the location of scores within a distribution. Some of the following measures of variability will contain more accurate estimates.
2. From a theoretical standpoint, since signs were ignored, this makes further mathematical manipulation of little value.

<u>VARIANCE</u>

Another measure of data distribution is the variance which is represented in a sample by the symbol s^2 (in a population by the symbol σ^2). The variance is calculated by squaring the deviation from the mean, totaling these values and then dividing them by the number of observations. The squaring operation eliminates the negative signs of the deviation and thereby avoids the absolute value roadblock that was encountered in the calculation of the mean deviation. The calculation of the variance on the previously used ham data would be as follows:

Ham No.	Ham Weight X	Ham Mean $\overline{X} = \frac{\Sigma X}{n}$	Deviation from the mean $x = X - \overline{X}$	Deviation Squared x^2 or $(X-\overline{X})^2$	Ham Weights Squared X^2
1	12	13.6	-1.6	+2.56	144
2	16	13.6	+2.4	+5.76	256
3	14	13.6	+0.4	+0.16	196
4	12	13.6	-1.6	+2.56	144
5	14	13.6	+0.4	+0.16	196
$\sum n=5$	$\sum X= 68$		$\sum x = 0$	$\sum x^2 = 11.20$	$\sum X^2=936$

$$s^2 = \frac{\sum (X-\bar{X})^2}{n} = \frac{\sum x^2}{n} = \frac{\sum f(X-\bar{X})^2}{n}$$

f = 1 if data are not in a frequency distribution table

$$s^2 = \frac{11.20}{5} = 2.24$$

Variance can often be calculated much easier from the raw data without figuring the deviation as follows:

$$s^2 = \frac{\sum X^2}{n} - \bar{X}^2 = \frac{\sum fX^2}{n} - \bar{X}^2$$

$$s^2 = \frac{936}{5} - (13.6)^2 = 187.2 - 184.96 = 2.24$$

STANDARD DEVIATION

Standard deviation is another measure of variability that is closely related to the variance and is defined as the square root of the variance. It is represented in the sample by the symbol s (in a population by the symbol σ). Its calculation is simply extracting the square root of the variance.

$$s = \sqrt{\frac{\sum (X-\bar{X})^2}{n}} = \sqrt{\frac{\sum X^2}{n} - \bar{X}^2} = \sqrt{s^2}$$

For the previous example:

$$s = \sqrt{2.24} = 1.4967$$

The standard deviation for discrete, or attribute as counting type variables is calculated as follows:

$$\text{Standard deviation of defectives} = \sqrt{\frac{\left(\text{Proportion of defective}\right) \times \left(\text{Proportion of good units}\right)}{\text{Number of sample units}}}$$

or

$$\text{Standard deviation as a \%} = \sqrt{\frac{(\%\ \text{defective}) \times (\%\ \text{good units})}{\text{Sample size}}}$$

Properties of the Variance and the Standard deviation

1. One of the properties of the mean, states that the sum of squares of the deviation from the mean is smaller than the sum of squares of the deviation from any other value. This means that the standard deviation, which uses in its calculation the sum of squares of the deviation from the mean, is also smaller than if any value other than the mean had been used.

2. For normal distributions the following relationship exists between the standard deviation and the percentage of the total observations that fall in different parts of the curve.

   a. Between the mean and + 1 s lies 34.13% of the total area under the curve.
      Between the mean and - 1 s lies 34.13% of the total area under the curve.
      Between - 1 s, through the mean, to + 1 s lies 68.27% of the total area under the curve.

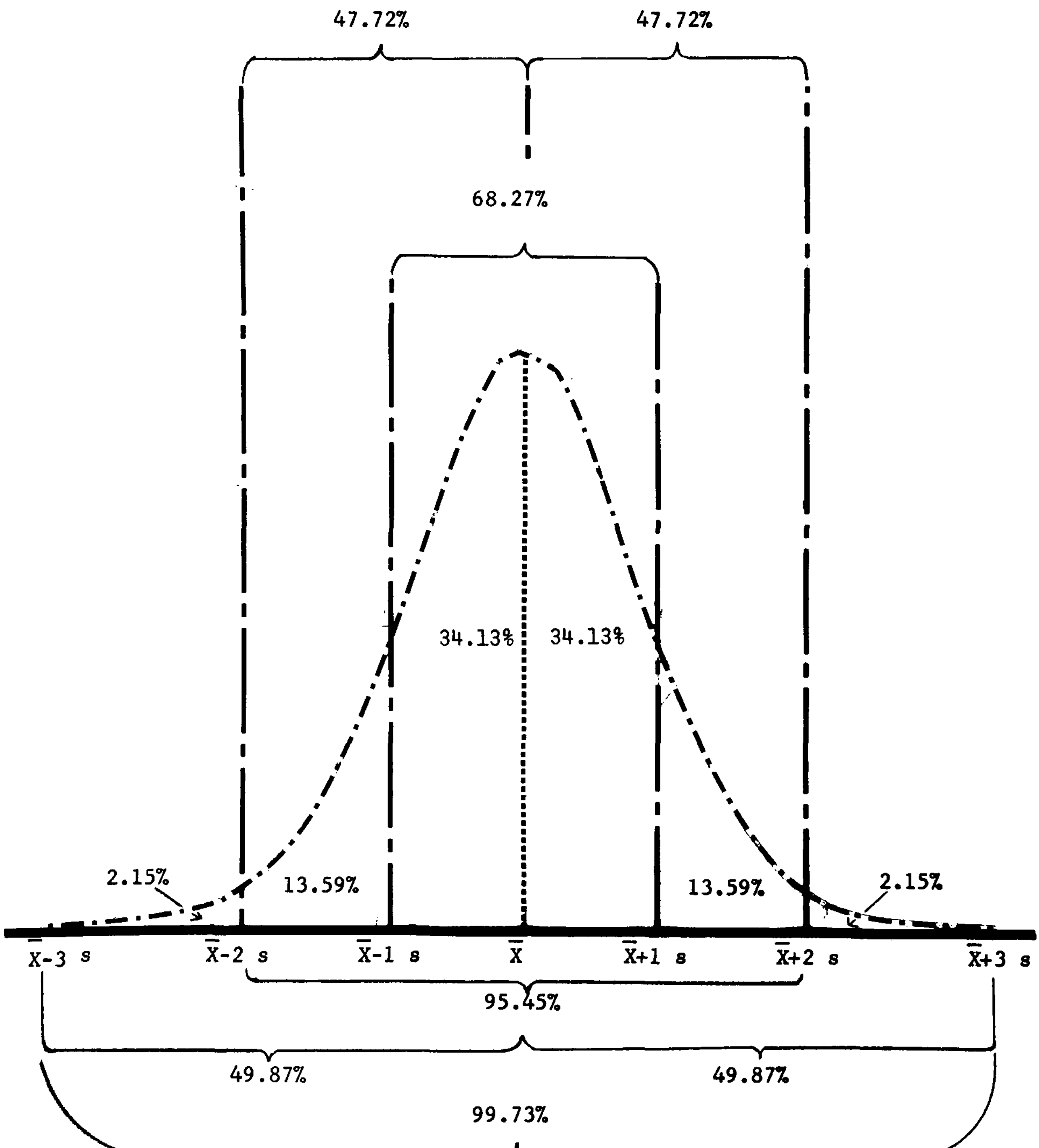
   b. Between + 1 s and + 2 s lies 13.59% of the total area under the curve.
      Between - 1 s and - 2 s lies 13.59% of the total area under the curve.
      Between the mean and + 2 s lies 47.72% of the total area under the curve.
      Between the mean and - 2 s lies 47.72% of the total area under the curve.
      Between - 2 s, through the mean, to + 2 s lies 95.45% of the total area under the curve.

   c. Between + 2 s and + 3 s lies 2.15% of the total area under the curve.
      Between - 2 s and - 3 s lies 2.15% of the total area under the curve.
      Between the mean and + 3 s lies 49.87% of the total area under the curve.

Between the mean and - 3 s lies 49.87% of the total area under the curve.

Between - 3 s, through the mean, to + 3 s lies 99.73% of the total area under the curve.

Graphically the normal curve can be subdivided as follows:



The example we have been using with only 5 observations is not large enough to approach a normal distribution. It will, however, be continued in order to illustrate that the distribution can vary considerably from the normal and fairly accurate estimates may still be obtained.

$$n = 5$$

$$\overline{X} = 13.6$$

$$s = 1.4967 = 1.5$$

Range	Lower limit	Upper limit	% of expected samples in this range	number of expected samples in this range	actual number of samples in this range
$\overline{X} \pm 1\ s$	13.6-1.5 = 12.1	13.6+1.5 = 15.1	68.27%	3.5	2
$\overline{X} \pm 2\ s$	13.6-2(1.5) = 10.6	13.6+2(1.5) = 16.6	95.45%	4.8	5
$\overline{X} \pm 3\ s$	13.6-3(1.5) = 9.1	13.6+3(1.5) = 18.1	99.73%	5.0	5

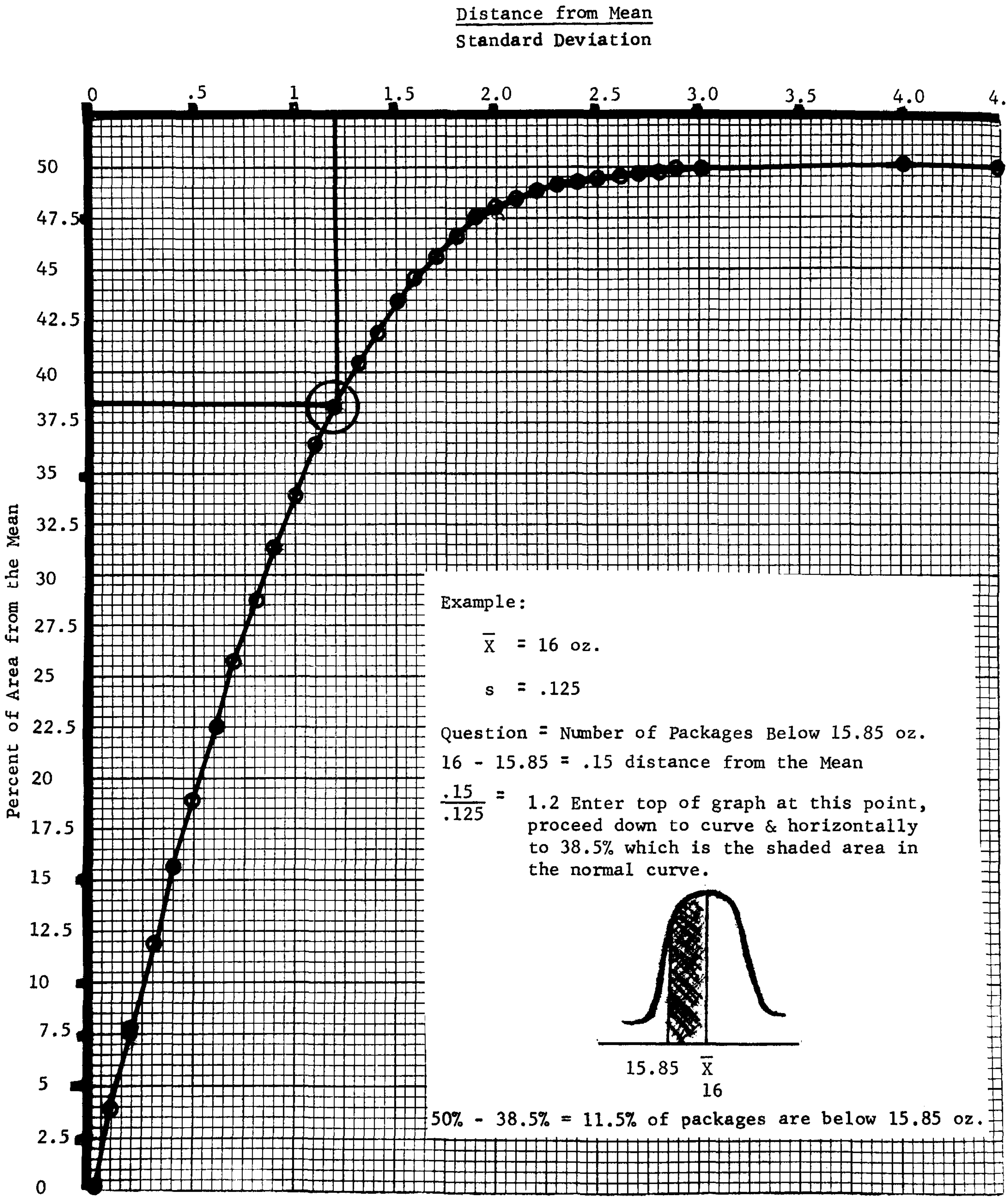
The last 2 columns show the comparison of expected distribution of observations with actual distribution.

3. The mean is often used as the best predictor for a population and the more the scores are concentrated around the mean (less variability) the better the prediction will be. The more variation in the data the poorer the prediction. Since the standard deviation is a measure of variability this means it is also an estimate of error (or precision) that is likely to be encountered when predictions are made.

   If predictions were made from two populations which have the same mean but different standard deviations, there would be more errors on the average from the sample with the largest standard deviation.

   If the standard deviation and the mean of a sample are known then the following graph will estimate the percent of the sample above or below any specific point in a sample.

AREA UNDER A NORMAL CURVE



Example:

$\overline{X}$ = 16 oz.

s = .125

Question = Number of Packages Below 15.85 oz.

16 - 15.85 = .15 distance from the Mean

$\frac{.15}{.125}$ = 1.2 Enter top of graph at this point, proceed down to curve & horizontally to 38.5% which is the shaded area in the normal curve.

50% - 38.5% = 11.5% of packages are below 15.85 oz.

COEFFICIENT OF VARIATION

Standard deviations may not be compared from one sample to the next unless:

1. The means are the same
2. The units of measurement are the same

For example, in estimating the live weight of animals and comparing it to the actual weight, a larger variation in the estimation errors would be expected with cattle than with hogs due to the fact that larger weights are involved. A 2% error, for example, would be 4 pounds in a 200-pound hog and 20 pounds in a 1000-pound steer.

Also, if our ham data had been expressed in ounces instead of pounds the following example would result:

Ham Number	Weight In pounds X_1	Weight In ounces X	X^2
1	12	192	36864
2	16	256	65536
3	14	224	50176
4	12	192	36864
5	14	224	50176
	$\Sigma X_1 = 68$	$\Sigma X = 1088$	$\Sigma X^2 = 239{,}616$

$\bar{X} = \frac{\Sigma X}{n}$

$\bar{X} = \frac{1088}{5}$

$\bar{X} = 217.6$

$$s = \sqrt{\frac{\Sigma X^2}{n} - \bar{X}^2} = \sqrt{\frac{239{,}616}{5} - (217.6)^2}$$

$$s = \sqrt{47{,}923.2 - 47{,}349.76} = \sqrt{573.44}$$

$$s = 23.95$$

When the weight is expressed in ounces instead of pounds the standard deviation is much greater than the original one calculated from the same data (1.4967).

To compare standard deviation from samples of different magnitudes it is necessary to calculate a coefficient of variation. The coefficient of variation is calculated by multiplying the standard deviation by 100 and dividing by the sample mean.

$$\text{Coefficient of variation} = \frac{100 \cdot s}{\bar{X}}$$

The results are expressed as a percent and this allows comparison of the variation in unrelated samples. Such factors as the variation in cattle carcass weights and the variation in hog carcass lengths can be compared to determine which has the largest variation.

The coefficient of variation for our ham problem would be as follows:

1. In pounds:

$$\text{Coefficient of variation} = \frac{100\ (1.4967)}{13.6} = 11.01\%$$

2. In ounces:

$$\text{Coefficient of variation} = \frac{100\ (23.95)}{217.6} = 11.01\%$$

These results indicate that the variation in the two examples are the same as would be expected.

ESTIMATION OF STANDARD DEVIATION FROM RANGE

Due to ease of calculation the range is often obtained rather than the standard deviation. But since the standard deviation is more useful in describing a population, a technique is needed to estimate the standard deviation from the range. The first problem encountered is the greater dependence of the range on the sample size and this is minimized by dividing the range by a divisor (d_2) obtained from the following graph. The d_2 factor varies with sample size and is equal to the average range ($\bar{R}$) divided by the population standard deviation.

$$d_2 = \frac{\bar{R}}{\sigma} \qquad\qquad \sigma \cong \frac{\bar{R}}{d_2}$$

Enter the graph on the right margin at the sample size location and proceed horizontally to the left until the curved line is intersected then proceed vertically down the page to obtain the divisor.

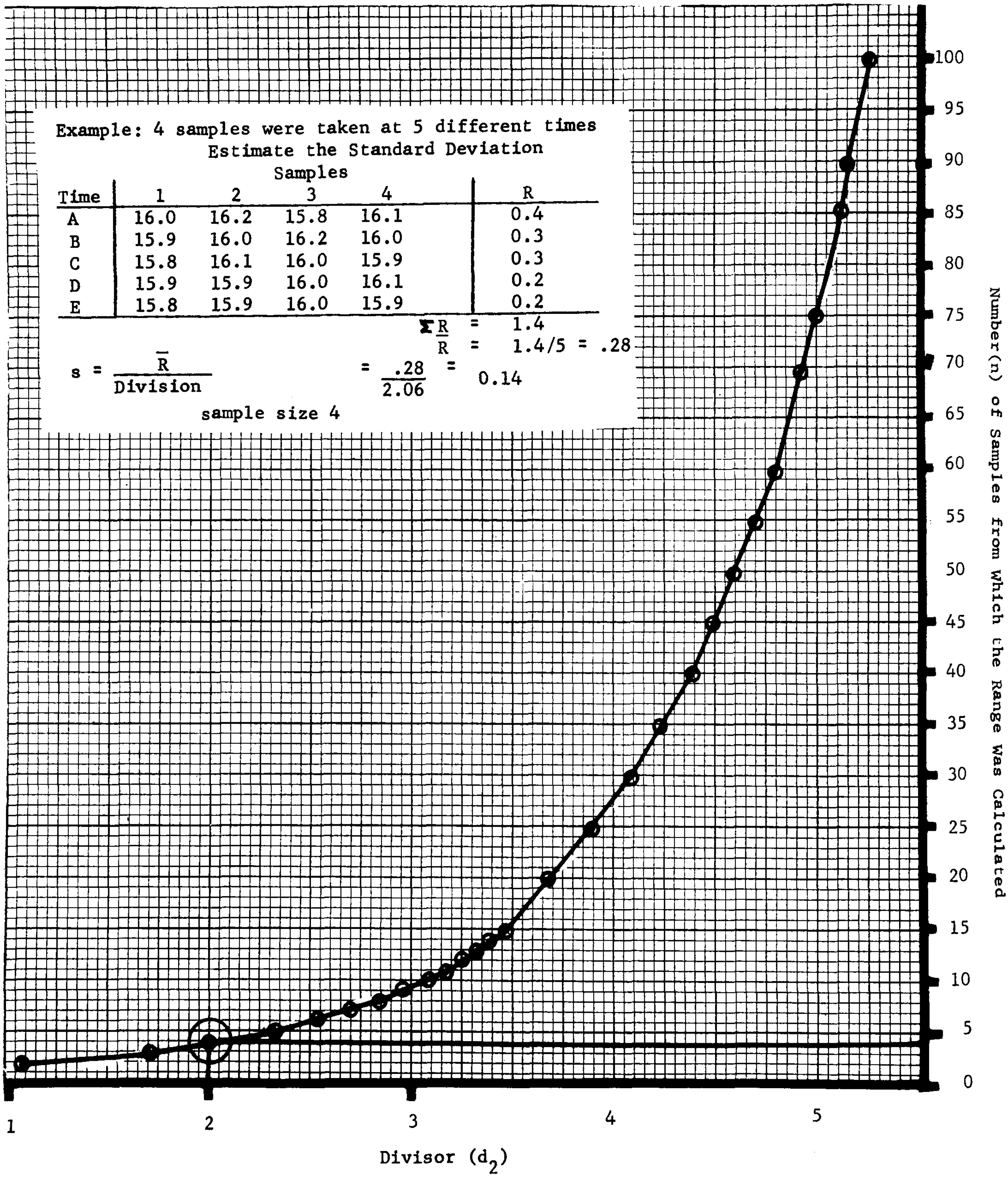
In the previous live weight of hogs example (this chapter, range) the range was 670 and the number of animals from which it was calculated was 6. Consulting the following graph a d_2 value of slightly greater than 2.5 is obtained. To estimate the standard deviation (s) of the hog weights the following procedure would be used.

$$\text{Standard deviation} = \frac{\text{range}}{\text{divisor}} = \frac{R}{d_2}$$

$$= \frac{670}{2.5} = 268$$

Probably this estimate of the standard deviation is slightly too high due to the one extreme observation of 850 lbs. If the standard deviation had been calculated by the normal method the results would be 240 pounds, indicating that this estimate from the range is slightly high.

ESTIMATING THE STANDARD DEVIATION FROM THE RANGE



Example: 4 samples were taken at 5 different times
Estimate the Standard Deviation

	Samples				
Time	1	2	3	4	R
A	16.0	16.2	15.8	16.1	0.4
B	15.9	16.0	16.2	16.0	0.3
C	15.8	16.1	16.0	15.9	0.3
D	15.9	15.9	16.0	16.1	0.2
E	15.8	15.9	16.0	15.9	0.2

$\Sigma R = 1.4$

$\bar{R} = 1.4/5 = .28$

$$s = \frac{\bar{R}}{\text{Division}} = \frac{.28}{2.06} = 0.14$$

sample size 4

Sample Size Influence on Measures of Central Tendency and Variability

Measures of Central Tendency:

In calculating the measures of central tendency usually the larger the sample the closer this estimate will be to the actual centers of the population. If the measures of central tendency are plotted for repeat samples from the same population the larger the samples are in size the smaller the variability (fluctuation) will be for these measures of central tendency. In order to prove this, a large population of frankfurters with a label weight of 16 oz. was repeatedly sampled for weight as follows:

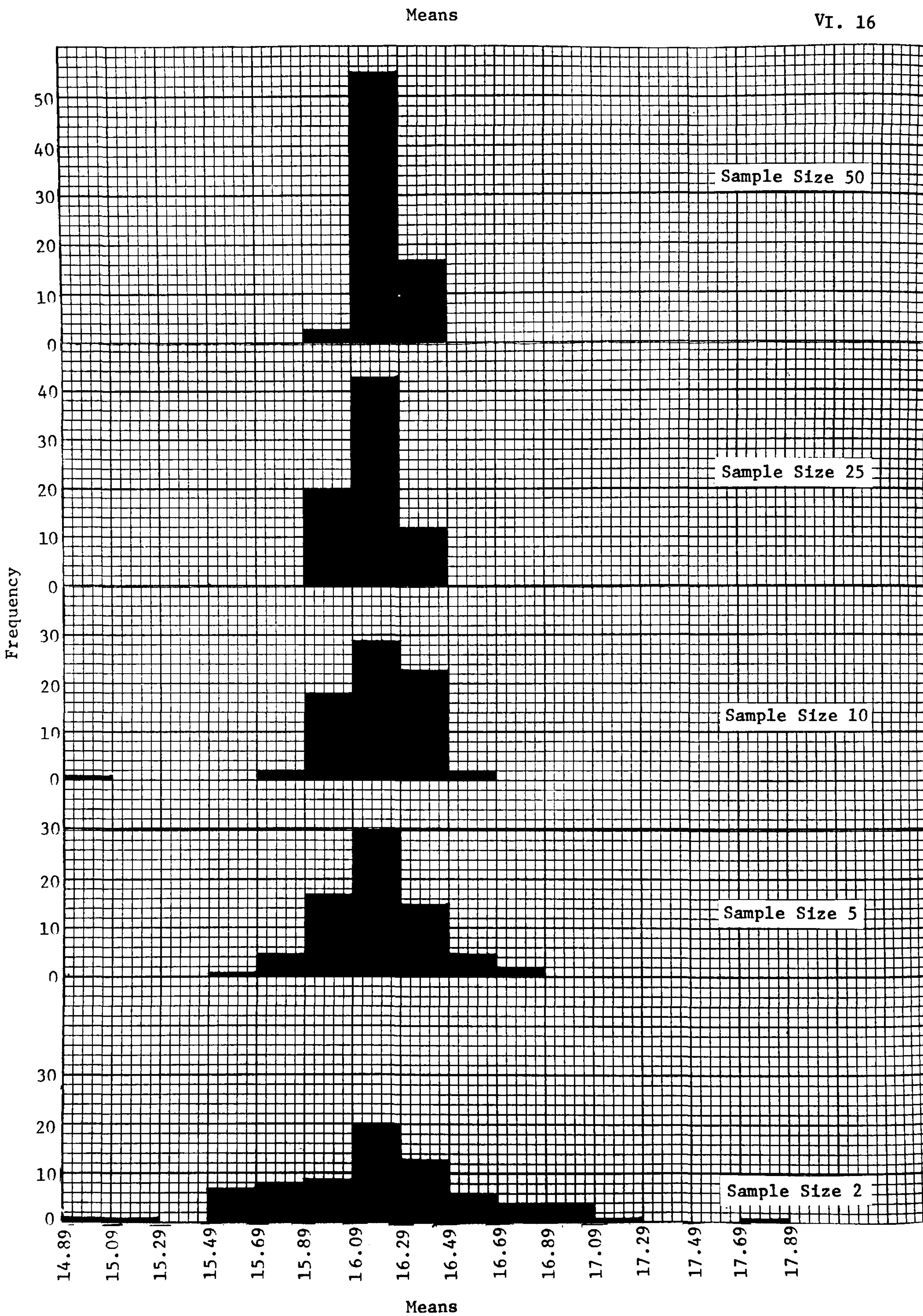
Number of Samples	Sample Size
50	50
50	25
50	10
50	5
50	2

These data were then plotted in the following 3 graphs (means, medians, modes). Notice that as a sample size increases from 2 to 50 the measures of central tendency are grouped closer together (less fluctuation in these measurements).

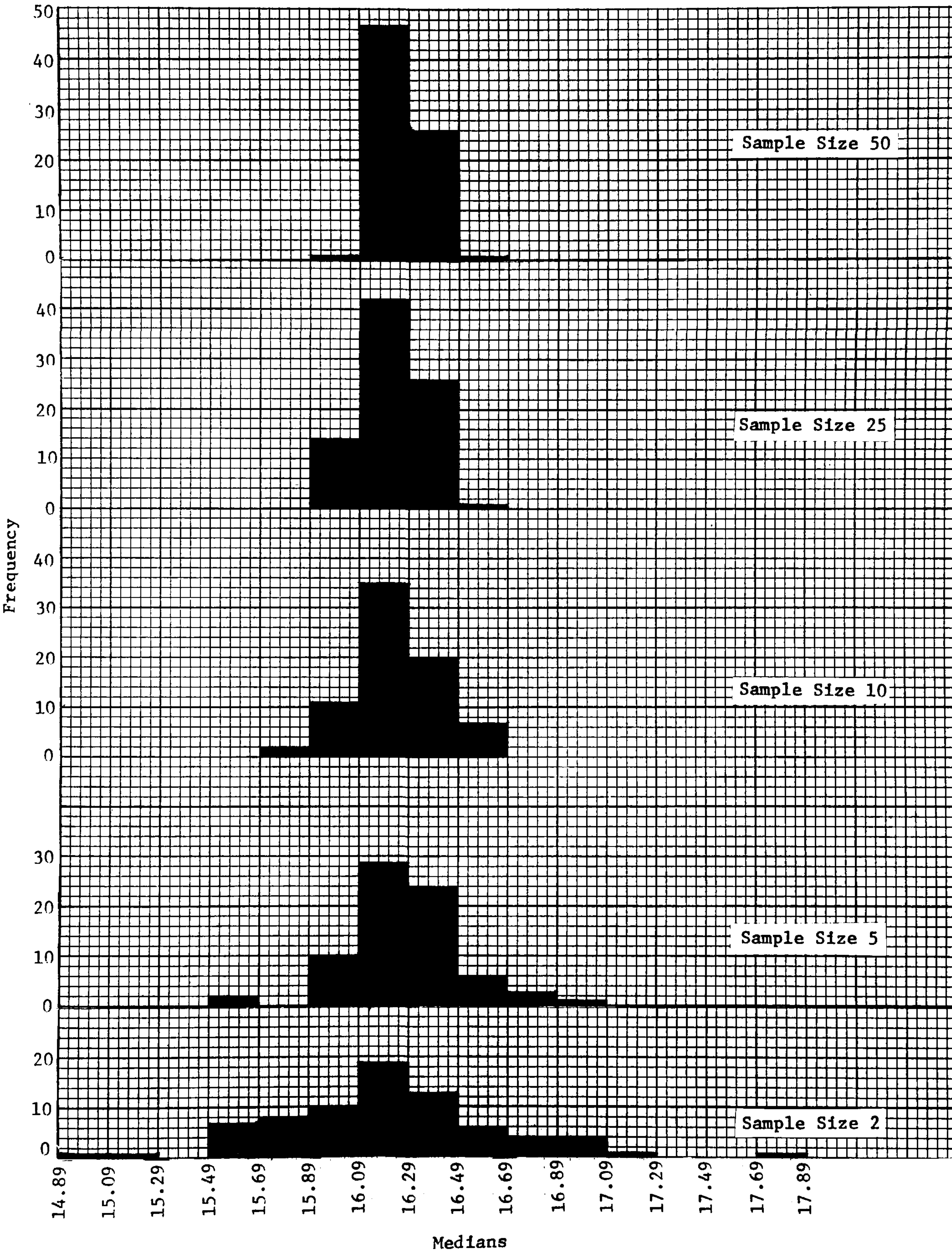
The mean had less variability on repeated samples than did the median, and the median had less variability than did the mode. For example the mean values fell in only 3 classes in the 25 and 50 samples size catagories but the median occupied 4 classes and the mode occupied 6 and 7 classes. This indicates that if a population is resampled the mean has less fluctuation between samples, (note 4th property of the mean in chapter on "Measures of Central Tendency").

Notice also that the class occupied most frequently (16.09-16.29) in all of the mean and median sample size catagories (2,5,10,25 and 50) is the same. This is also the same for the mode except for sample size 10. This would suggest that the mode is the least reliable measure of central tendency, (Note mode is described as a <u>rough</u> approximation of the center of the population in chapter on "Measures of Central Tendency").

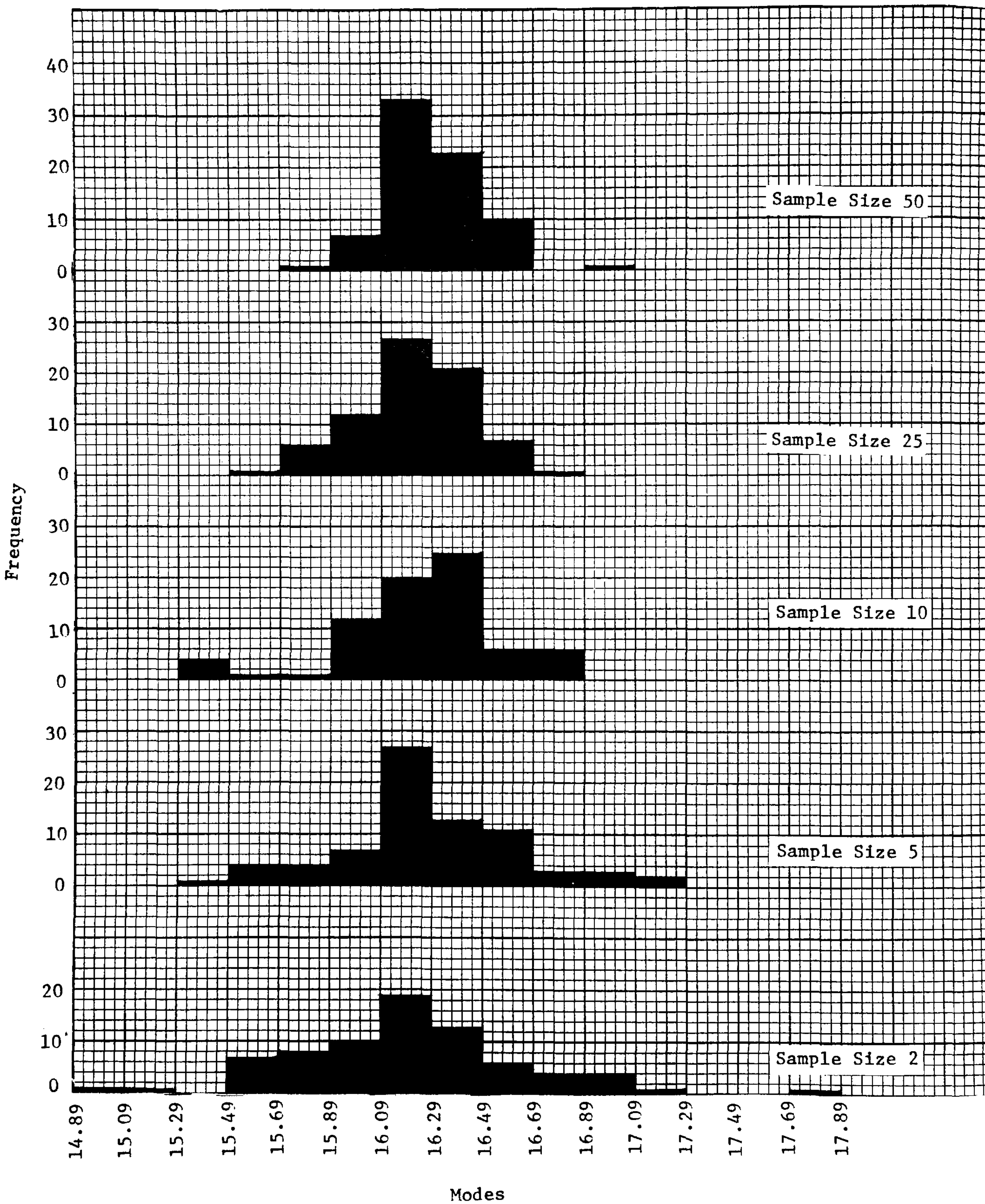
To summarize: (1) as sample size increases there is less fluctuation in measures of central tendency on repeated samples, (2) the mean is the most consistent measure of central tendency, (3) the mode is the least repeatable and reliable measure of central tendency.


Means
Frequency
Sample Size 50
Sample Size 25
Sample Size 10
Sample Size 5
Sample Size 2
0
10
20
30
40
50
14.89
15.09
15.29
15.49
15.69
15.89
16.09
16.29
16.49
16.69
16.89
17.09
17.29
17.49
17.69
17.89
Means

Medians


Frequency
Sample Size 50
Sample Size 25
Sample Size 10
Sample Size 5
Sample Size 2
50
40
30
20
10
0
14.89
15.09
15.29
15.49
15.69
15.89
16.09
16.29
16.49
16.69
16.89
17.09
17.29
17.49
17.69
17.89
Medians

Modes


40
30
20
10
0
Sample Size 50
30
20
10
0
Sample Size 25
Frequency
30
20
10
0
Sample Size 10
30
20
10
0
Sample Size 5
20
10
0
Sample Size 2
14.89
15.09
15.29
15.49
15.69
15.89
16.09
16.29
16.49
16.69
16.89
17.09
17.29
17.49
17.69
17.89
Modes

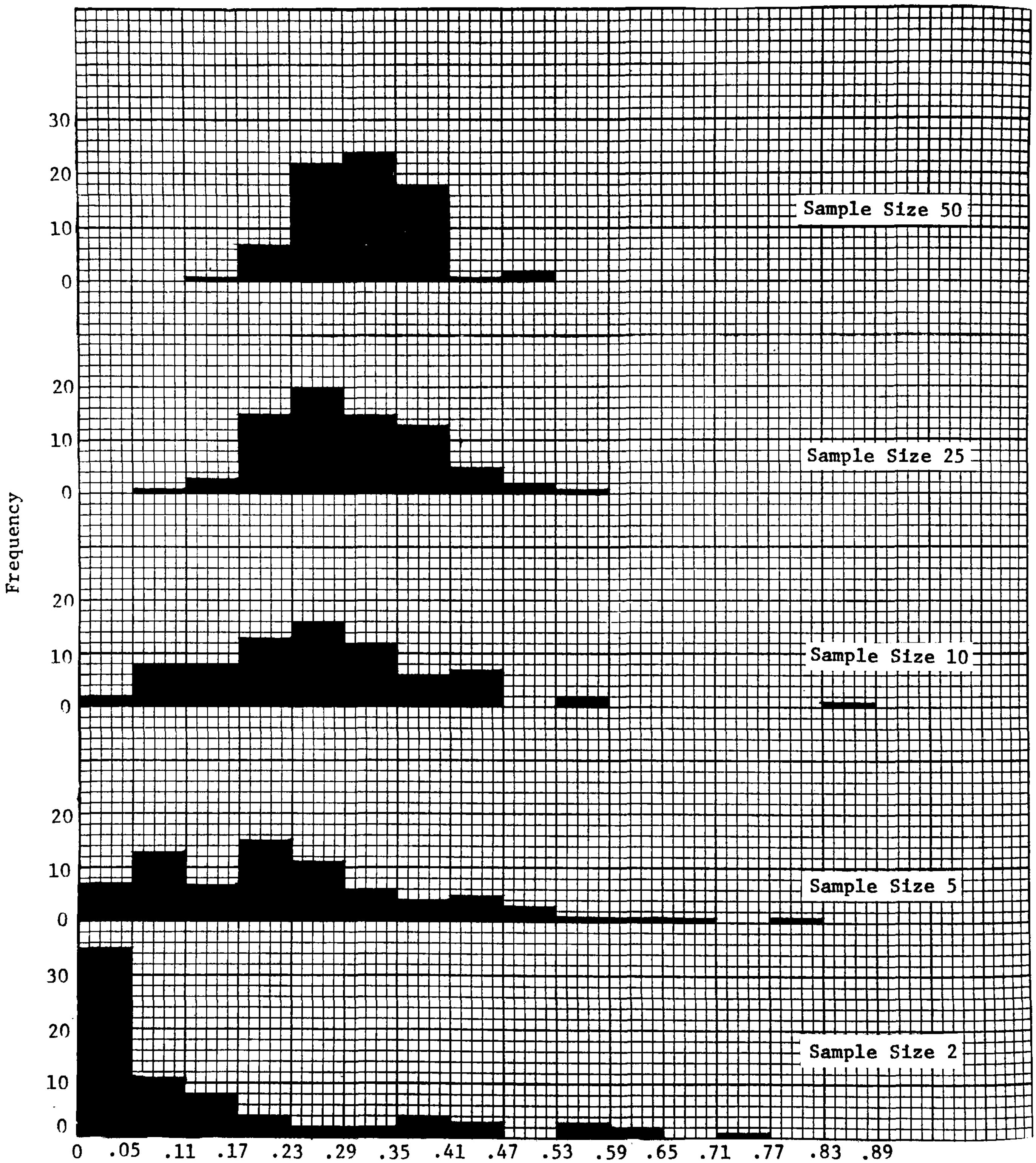
From these same data measures of variability such as the variance and standard deviation were calculated and are illustrated in the next two graphs (Variance and standard deviation $= \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}}$). From the variance graph it is obvious that this measure of variability is influenced by sample size and it is particularly noticeable in the small sample groups (sample size 2 and 5). Extracting the square root for the standard deviation renders this less obvious but it is still quite noticeable particularly in the small sample size group (sample size 2).

Many statisticians recommend an alteration in the standard deviation formula to help alleviate this situation. The formula recommended is:

$$\text{Standard deviation} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = \sqrt{\frac{\Sigma X^2 - n(\bar{X})^2}{n-1}} = \sqrt{\frac{n\Sigma X^2 - (\Sigma X)^2}{n(n-1)}}$$

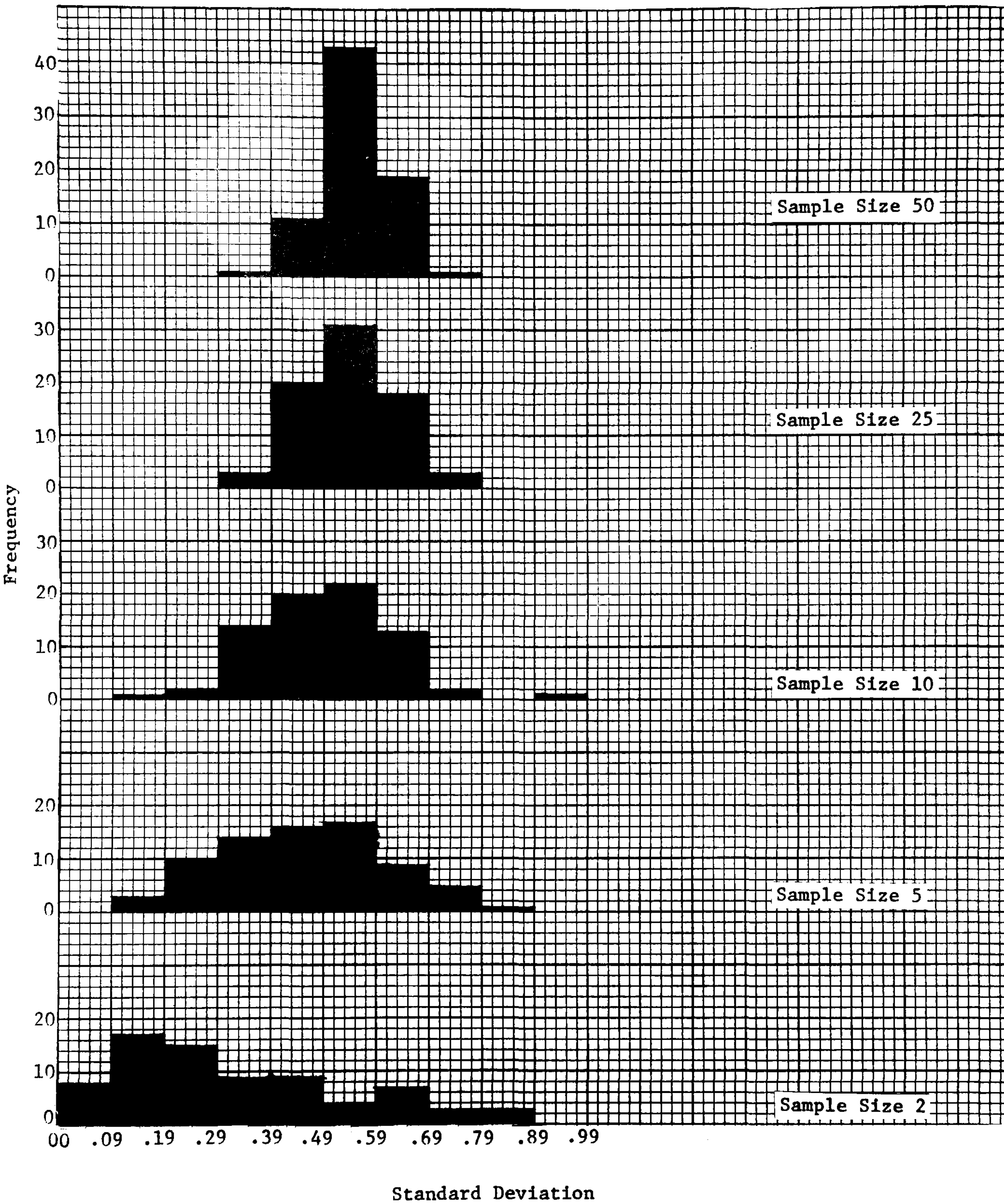
Using this formula a third graph (Standard deviation $= \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$) is plotted and proves to be more consistent with less fluctuation in the standard deviation measurement than the previous formula. There are still problems (although not as large a magnitude) with very small samples (sample size 2) but because of the less dependence on sample size the revised formula is often used (when used in this text it will be indicated).

Variance

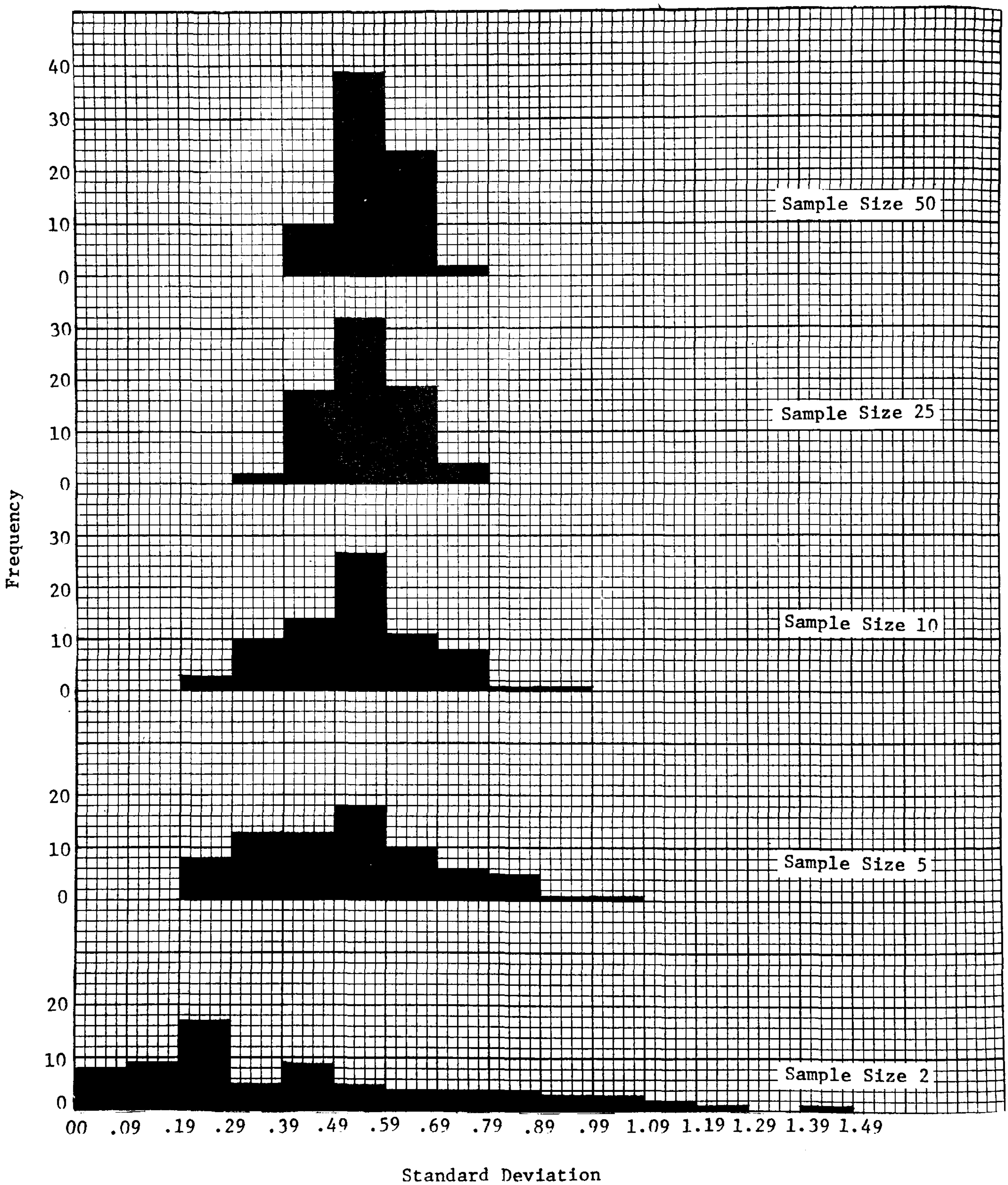


Variance

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}}$$

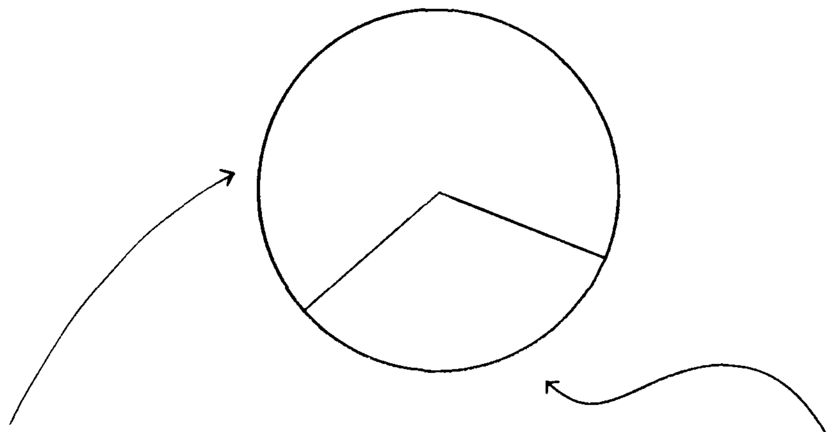


$$\text{Standard Deviation} = \sqrt{\frac{\Sigma\ (X-\overline{X})^2}{n-1}}$$



Total variation can be divided into 2 parts.

Total Variation



Chance or Random or natural variation	Assignable or correctable as unnatural variation
Is always present	Arrives from specific disorders of the process
It is inherent in the process	
Not identifiable	Identifiable
Not removable	Removable
Not within our power to regulate	Within our power to regulate
Predictable within limits	Merits prompt investigation

SAMPLE PROBLEMS

1. Range - measure of variability. It is the low value subtracted from the high value.

Example:

Observations

X	Range = Big-Small
9	
7	= 9 - 1 = 8
1	
2	= or 1 to 9

Calculate the range of the following group of samples.

A	B	C	D	E	F	G	H	I	J
9	1	3	100	3	1	8	1	2	6
10	9	2	2	3	2	0	1	2	8
5	3	7	9	3		9	1	2	4
	6	3	27	3		3			3
		4	31	1					
		5	19	7					
		5	6	3					
			24	3					
			9						
			7						
			6						
			6						
			6						
			6						

Name an advantage and a disadvantage of the range.

Answers (A-5, B-8, C-5, D-98, E-6, F-1, G-9, H-0, I-0, J-5)

2. Mean Deviation - a measure of variability. Steps to calculate.

   1. Calculate the mean ($\overline{X}$).
   2. Calculate the deviation from the mean ($X-\overline{X}$) and check the accuracy of the mean. Note the difference in the meaning of the terms "mean deviation" and "deviation from the mean".
   3. Total the deviation without regard to sign.
   4. Divide by the number of observations.

Example:

Observation X	Mean $\overline{X}$	Deviation $(X-\overline{X})=x$
2	3	-1
3	3	0
6	3	+3
2	3	-1
2	3	-1

$\sum X=15$

$\overline{X}=\frac{\sum X}{n}$

$=\frac{15}{5}=3$

$\sum(X-\overline{X})=0$ - mean is correct

$\sum|X-\overline{X}|=6$

mean deviation $=\frac{\sum|X-\overline{X}|}{n}$

$=\frac{6}{5}$

$= 1.2$

Calculate the mean deviation for the following group of samples.

A	B	C	D	E	F	G	H	I	J
2	4	3	6	5	2	4	1	3	9
3	2	3	2	1	9	6	1	1	1
4	0	3	1	3	8	5	2	7	8
	2		7	6	6		2	1	2
				8				3	
				7					

Name 2 properties of the mean deviation.

Answers (A-.66, B-1.0, C-0, D-2.5, E-2, F-2.25, G-.66, H-.5, I-1.6, J-3.5)

3. Variance - a measure of variability. Steps in calculation.

1. Calculate the mean $(\bar{X})$.
2. Calculate the deviations from the mean $(X-\bar{X})$. (Check the accuracy of the mean).
3. Square the deviations from the mean $(X-\bar{X})^2$.
4. Add the deviations squared $\left[\sum (X-X)^2\right]$.
5. Divide the sum of deviations squared by the number of observations. $\dfrac{\sum (X-\bar{X})^2}{n}$

Example:

Observation X	Mean $\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
4	5	-1	1
3	5	-2	4
5	5	0	0
8	5	+3	9
$\sum X=20$		$\sum (X-X)=0$	$\sum (X-\bar{X})^2=14$

$$\mathrm{Var}=(s^2) = \frac{\sum (X-\bar{X})^2}{n}$$

$$= \frac{14}{4} = 3.5$$

↑ Mean is correct

$$\bar{X}= \frac{\sum X}{n} = \frac{20}{4} = 5$$

Variance check

Steps in calculation
1. Square the original observation X^2
2. Total the squared observations $\sum X^2$
3. Divide the sum of squares by the number of observations $\dfrac{\sum X^2}{n}$
4. Square the mean value $\bar{X}^2$
5. Subtract number 4 from number 3 or subtract the mean squared from the sum of squares divided by the number of observations. $\dfrac{\sum X^2}{n} - \bar{X}^2$
6. If the variance calculation is correct this value should be the same as the variance calculated by the original procedure.

Example of checking the variance:

Original observations X	Original observations squared X^2
4	16
3	9
5	25
8	64
$\sum X=20$	$\sum X^2=114$

$$\bar{X}= \frac{\sum X}{n}$$

$$= \frac{20}{4}$$

$$= 5$$

$$\mathrm{Var}=(s2) = \frac{\sum X^2}{n} - \bar{X}^2$$

$$\mathrm{Var} = \frac{114}{4} - (5)^2$$

$$= 28.5 - 25$$

$$= 3.5$$ ⟵ check with original variance

Variance

Calculate the variance of the following groups and check your calculation for accuracy.

A	B	C	D	E	F	G	H	I	J
6	7	4	3	4	2	6	2	4	1
2	3	0	3	3	1	4	4	9	8
1	5	2	3	2	3	5	6	0	1
	5	2			5	5		3	0
					4	3			
						7			

Answer (A-4.67, B-2, C-2, D-0, E-.67, F-2, G-1.67, H-2.67, I-10.5, J-10.25)

4. Standard Deviation - measure of variability. It is calculated by taking the square root of the variance (s^2).

Example:

Observations

X
4
3
5
8

$\sum X = 20$

$$\bar{X} = \frac{\sum X}{n} = \frac{20}{4} = 5$$

$$Var(s^2) = \frac{\sum (X-\bar{X})^2}{n} = \frac{14}{4} = 3.5$$

$$s = \sqrt{s^2} = \sqrt{3.5} = 1.9$$

Check on square root calculation

$$\begin{array}{r} 1.9 \\ \underline{1.9} \\ \underline{17.1} \\ \underline{19} \\ 3.61 \\ \underline{-\ 11} \\ 3.50 \end{array}$$

Calculate the standard deviation of the following group of samples. The examples on the "Variance Problem" will be used to calculate the standard deviation.

	A	B	C	D	E	F	G	H	I	J
Variance	4.67	2	2	0	.67	2	1.67	2.67	10.5	10.25

What percent of the sample observation from the variance problem fall within the range of the mean $\pm$ one standard deviation.

On the average what percent is expected within this range?

Answers (A-2.16, B-1.4, C-1.4, D-0, E-.82, F-1.4, G-1.29, H-1.63, I-3.24, J-3.20)

5. Coefficient of variation is a measure of variability and is calculated by multiplying the standard deviation(s) by 100 and divided by the mean ($\bar{X}$).

$$CV = \frac{s.100}{\bar{X}}$$

Continuing the example from the "Variance" and "Standard Deviation" problems.

Example:

Observations X
4
3
5
8

$\sum X = 20$

Variance = 3.5

Standard Deviation = 1.9

$$\bar{X} = \frac{\sum X}{n}$$

$$= \frac{20}{4}$$

$$= 5$$

$$CV = \frac{1.9 \times 100}{5}$$

$$= 38\%$$

Calculate the coefficient of the following groups that were continued from the variance and standard deviation examples.

	A	B	C	D	E	F	G	H	I	J
Mean	3	5	2	3	3	3	5	4	4	2.5
Variance	4.67	2	2	0	.67	2	1.67	2.67	10.5	10.25
Standard Deviation	2.16	1.4	1.4	0	.82	1.4	1.29	1.63	3.24	3.20

Which sample had the most variation?

Which sample had the least variation?

How can the coefficient of variation be used?

Answers (A-72%, B-28%, C-70%, D-0%, E-27.33%, F-46.67%, G-25.8%, H-40.75%,
I-81%, J-128%)

6. The following five samples were observed.

A	4
B	6
C	5
D	4
E	6

A. Calculate the mean and check the answer.

B. Calculate the mode.

C. Calculate the median.

D. Calculate the mean deviation.

E. Calculate the range.

F. Calculate the variance and check the answer.

G. Calculate the standard deviation and check the square root calculation.

H. Calculate the coefficient of variation.

Answers (A-5, B-5, C-5, D-.8, E-2, F-.8, G-.89, H-17.8%)

7. The following five samples were observed.

A	4
B	3
C	6
D	7
E	10

A. Calculate the mean and check the answer.

B. Calculate the mode.

C. Calculate the median.

D. Calculate the mean deviation.

E. Calculate the range.

F. Calculate the variance and check the answer.

G. Calculate the standard deviation and check the square root calculation.

H. Calculate the coefficient of variation.

Answers (A-6, B-No mode, C-6, D-2, E-7, F-6, G-2.45, H-40.8%)

References

Haber, Audrey and Richard P. Runyon. 1969. General Statistics. Addison-Wesley Publishing Company, Reading, Massachusetts; Menlo Park, California; Don Mills, Ontario; London.

Kramer, A. and B. A. Twigg. 1970. Quality Control for the Food Industry. Vol. I. The AVI Publishing Co., Inc., Westport, Conn.

Smith, G. Milton. 1962. A Simplified Guide to Statistics. Holt, Rinehart and Winston, Inc., New York.

Spiegel, Murray R. 1961. Theory and Problems of Statistics. McGraw-Hill Book Company, New York, St. Louis, San Francisco, Toronto, Sydney.

Walker, Helen M. and Joseph Lev. 1958. Elementary Statistical Methods. Holt, Rinehart and Winston, New York.

PROBABILITY

Probability is the likelihood of a given event occurring out of the total possible events that could occur. If the event occurs it is called a success and the probability of a success is illustrated by the following formula:

$$p = \frac{h}{n}$$

p = probability of a success
h = ways in which a success may be obtained
n = total possible, equal likely ways that the event could occur

Example: There are 3 bad wieners in a package of 12. By randomly selecting one wiener, what is the probability of selecting a bad one (called a success)?

$$p = \frac{3}{12} = 1/4$$

The probability of the event not occurring is called a failure and is illustrated by the following formula:

$$q = \frac{n-h}{n} = 1-p$$

q = probability of failure

In our previous example:

$$q = \frac{12-3}{12} = \frac{9}{12} = \frac{3}{4}$$

or

$$q = 1 - \frac{1}{4} = \frac{3}{4}$$

Properties of the probability:

1. A probability of zero means an event cannot occur
2. Probability of one means an event is certain to occur

3. $p + q = 1$

4. Odds constructed from probability are p to q. In our example the odds are 1/4 to 3/4 or 1 to 3 of selecting one of the bad wieners with one random drawing.

Independent events:

If the success or failure of one trial has no effect on a subsequent trial the events are said to be independent. If the events are independent, the probability of all trials yielding a success may be found by multiplying the probability of each trial involved.

Example: A wiener is randomly drawn from our previous example, examined and replaced in the package. A second wiener is randomly drawn from the package of 12. The drawing and replacing of the 1st wiener has no effect on the outcome of the drawing of the second wiener and, therefore, these are independent events. What is the probability of both (1st and 2nd drawing) wieners being bad?

$$p = p(A) \cdot p(B)$$

$$p = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$$

Dependent events:

If the success or failure of one trial has an effect on a subsequent trial the events are said to be dependent. In this case the probability of a latter trial must be calculated assuming the results of an earlier event.

Example: A wiener is randomly drawn from our previous example and examined. It is not replaced in the package. A second wiener is drawn from the remaining eleven. What is the probability of both (1st and 2nd drawing) of these being bad?

Probability of 1st trial has not changed.

$$p = p(A)$$

$$p = \frac{3}{12} = \frac{1}{4}$$

To obtain a 2nd bad wiener out of 2 trials the 1st one drawn has to be a bad wiener or the total (drawing 2 bad wieners) is a failure. If the 1st one drawn is a bad wiener and not replaced the remainder of the sample would contain 11 wieners of which 2 are bad. This makes the probability on the 2nd drawing as follows:

$p + p(B \mid A)$ reads probability of B given A

$$p = \frac{2}{11}$$

The total probability of both being bad thus becomes:

$$p = p(A) \cdot p(B \mid A)$$

$$p = \frac{1}{4} \cdot \frac{2}{11} = \frac{2}{44} = \frac{1}{22}$$

<u>Mutually exclusive</u>:

If one event occurring means that other events cannot occur the events are said to be mutually exclusive. In mutually exclusive events the probability of <u>either of the events</u> occurring is the sum of their individual probabilities.

Example: In our package of 12 wieners the 3 bad ones were caused by bacterial growth due to a lack of salt concentration. Two other wieners in the same package were oxidized due to an excess of salt concentration.

Under these conditions a sample cannot be bad and oxidized both. (Note, this applies only to this example). This means that the events are mutually exclusive. The probability of obtaining in one random drawing a bad or an oxidized wiener is illustrated as follows:

Bad wiener: $p = p(A)$

$$p = \frac{3}{12} = \frac{1}{4}$$

Oxidized wiener: $p = p(B)$

$$p = \frac{2}{12} = \frac{1}{6}$$

Bad or oxidized (mutually exclusive)

$$p = p(A) + p(B)$$

$$p = \frac{1}{4} + \frac{1}{6} = \frac{3}{12} + \frac{2}{12} = \frac{5}{12}$$

Not Mutually exclusive:

If one event occurring does not exclude another event from occurring the events are not mutually exclusive. In non-mutually exclusive events the probability of either of the events occurring or of both occurring (same as probability of at least one occurring) may be found by adding the probability of the individual events occurring and subtracting the probability of the non-mutually exclusive events occurring.

Example: In our previous example if one of the bad wieners had also been oxidized then we could have drawn a bad oxidized wiener. We now have in our sample of 12 wieners the following:

2 - bad wieners

1 - bad and oxidized wiener

2 - oxidized wieners

7 - good wieners

Then the probability of drawing a bad wiener or an oxidized wiener or one which is both bad and oxidized is:

$$p = p(A) + p(B) - p(AB)$$

$$p = \underbrace{\frac{3}{12}}_{\text{bad}} + \underbrace{\frac{3}{12}}_{\text{oxidized}} - \underbrace{\frac{1}{12}}_{\text{bad \& oxidized}} = \frac{5}{12}$$

To summarize the probability formulas:

General Probability:

$$p = \frac{h}{n}$$

$$q = \frac{n-h}{n}$$

Independent Events (more than one sample selected): (Probability of both occurring)

$$p = p(A) \cdot p(B)$$

Dependent Events (more than one sample selected): (Probability of both occurring)

$$p = p(A) \cdot p(B|A)$$

Mutually Exclusive (one sample selected): (Probability of either occurring)

$$p = p(A) + p(B)$$

Not Mutually Exclusive (one sample selected): (Probability of at least one occurring)

$$p = p(A) + p(B) - p(AB)$$

Sample Problems

What is the probability of the following -

1. A 6 on the toss of one die?

2. An Ace drawn from a deck (52) of cards?

3. Of not drawing a heart from a deck of cards?

Cards are drawn and replaced in a deck (52) -

4. In two drawings, what is the probability of both being hearts?

5. On two drawings, what is the probability of both being aces?

Cards are drawn and not replaced in deck -

6. On two drawings, what is the probability of both being kings?

7. On two drawings, what is the probability of both being spades?

The Ace of Hearts has been discarded from a deck -

8. What is the probability of drawing a heart?

9. What is the probability of drawing either and ace or a heart?

Full deck -

10. What is the probability of drawing an ace or a heart?

(Answers - 1-1/6, 2-1/13, 3-39/52, 4-1/16, 5-1/169, 6-1/221, 7-1/17, 8-4/17, 9-5/17, 10-4/13)

References

Anderson, R. L. and T. A. Bancroft. 1952. Statistical Theory in Research. McGraw-Hill Book Company, Inc., New York, Toronto, London.

Berman, Simeon J. 1969. The Elements of Probability. Addison-Wesley Publishing Company, Reading, Massachusetts; Menlo Park, California; London; Don Mills, Ontario.

McCarthy, Philip J. 1957. Introduction to Statistical Reasoning. McGraw-Hill Book Company, Inc., New York, Toronto, London.

Mood, Alexander McFarlane. 1950. McGraw-Hill Book Company, Inc., New York, Toronto, London.

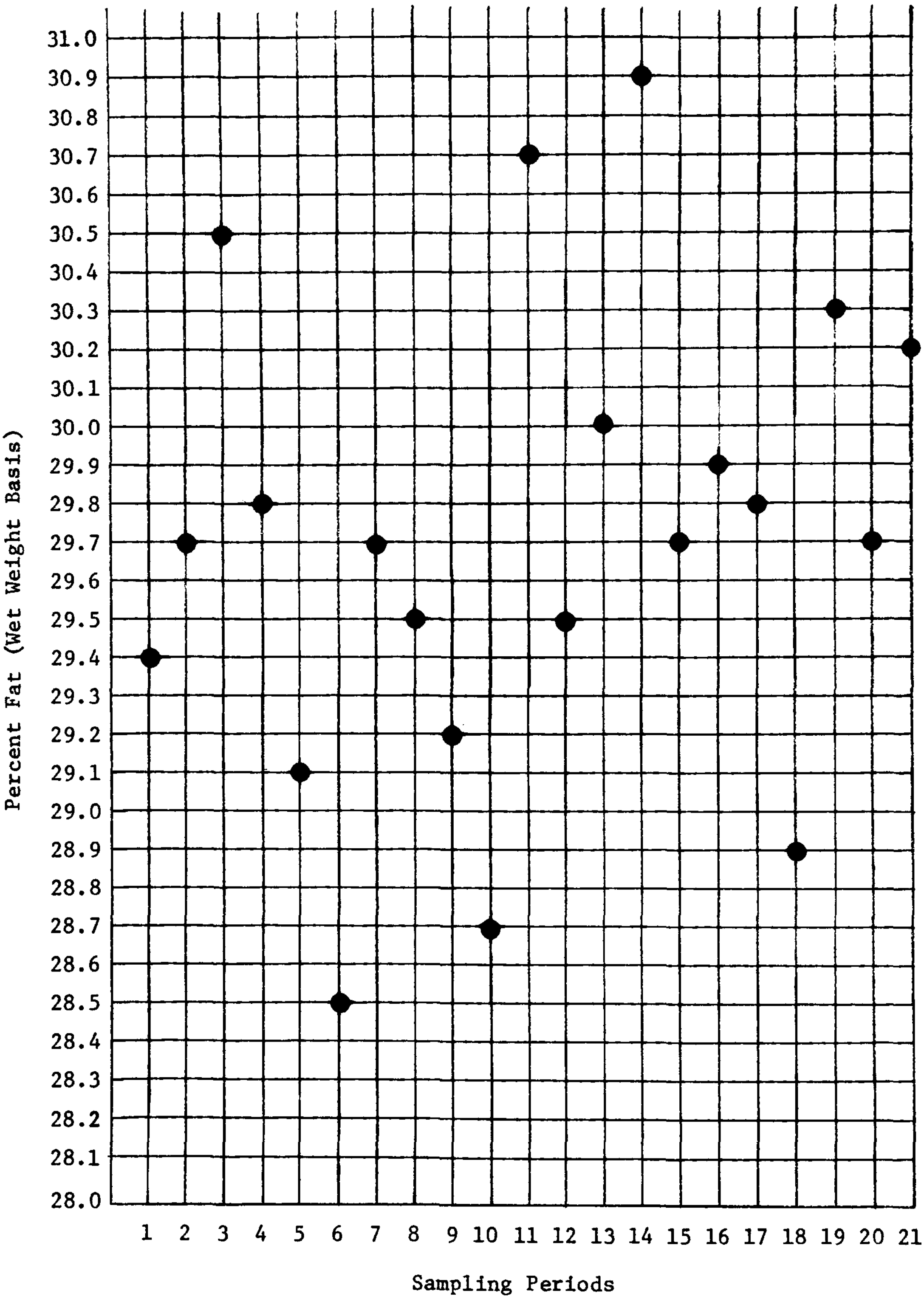
Spiegel, Murray R. 1961. Theory and Problems of Statistics. McGraw-Hill Book Company, New York, St. Louis, San Francisco, Toronto, Sydney.

QUALITY CONTROL CHARTS

DESCRIPTION OF A POPULATION

A measure of central tendency and measure of variability are both needed to describe a population.

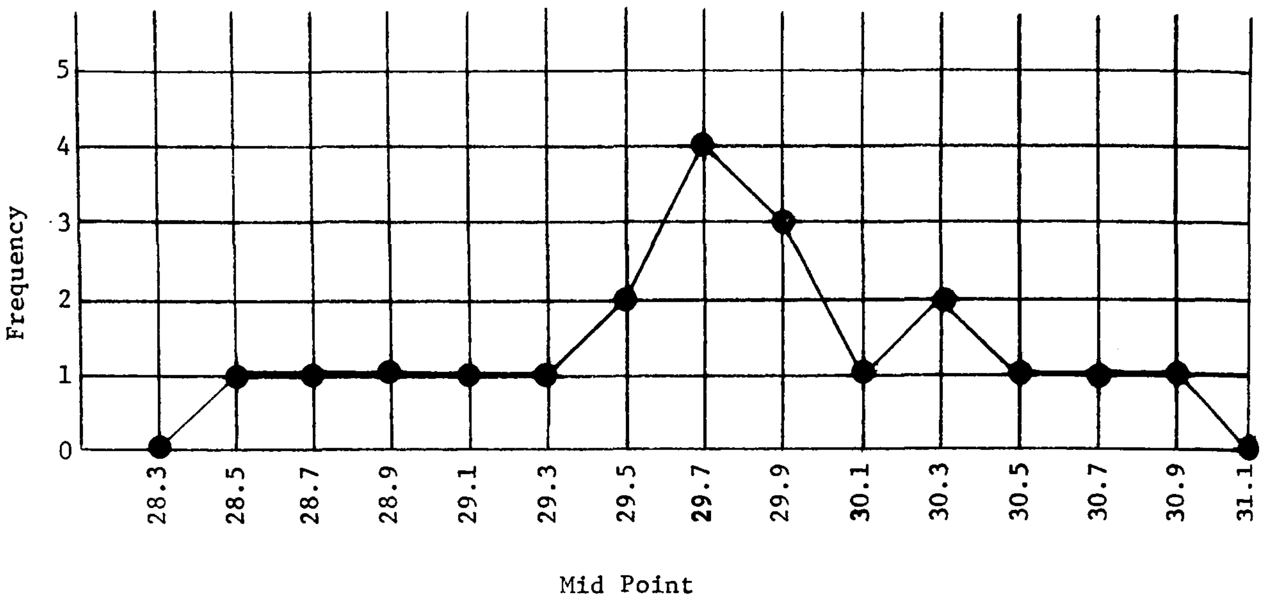
An example of how this information might be applied in a meat quality control program is shown by the following graph which plots the percent chemical fat found on repeated samplings of a frankfurter product.



If a frequency table was constructed from this information it would look as follows:

Lower Limits	Upper Limits	Mid Point		Frequency
28.20	28.39	28.3		0
28.40	28.59	28.5		1
28.60	28.79	28.7		1
28.80	28.99	28.9		1
29.00	29.19	29.1		1
29.20	29.39	29.3		1
29.40	29.59	29.5		3
29.60	29.79	29.7		4
29.80	29.99	29.9		3
30.00	30.19	30.1		1
30.20	30.39	30.3		2
30.40	30.59	30.5		1
30.60	30.79	30.7		1
30.80	30.99	30.9		1
31.00	31.19	31.1		0

Plotting this percent fat information on a graph would give the following picture.



A graph has now been constructed of the population and the next task becomes one of describing this population and this requires a measure of central tendency and a measure of variability. In this problem the mean and the standard deviation will be used.

X	$\bar{X}$	$x = (X-\bar{X})$	$x^2 = (X-\bar{X})^2$	X^2
29.4	29.7	-.3	.09	864.36
29.7	29.7	0	0	882.09
30.5	29.7	+.8	.64	930.25
29.8	29.7	+.1	.01	888.04
29.1	29.7	-.6	.36	846.81
28.5	29.7	-1.2	1.44	812.25
29.7	29.7	0	0	882.09
29.5	29.7	-.2	.04	870.25
29.2	29.7	-.5	.25	852.64
28.7	29.7	-1.0	1.00	823.69
30.7	29.7	+1.0	1.00	942.49
29.5	29.7	-.2	.04	870.25
30.0	29.7	+.3	.09	900.00
30.9	29.7	+1.2	1.44	954.81
29.7	29.7	0	0	882.09
29.9	29.7	+.2	.04	894.01
29.8	29.7	+.1	.01	888.04
28.9	29.7	-.8	.64	835.21
30.3	29.7	+.6	.36	918.09
29.7	29.7	0	0	882.09
30.2	29.7	+.5	.25	912.04
$\sum X = 623.7$		$\sum(X-\bar{X}) = 0$	$\sum(X-\bar{X})^2 = 7.70$	$\sum X^2 = 18{,}531.59$
		$\sum x = 0$	$\sum x^2 = 7.70$	

$$\bar{X} = \frac{\sum X}{n} = \frac{623.7}{21} = 29.7$$

$$s^2 = \frac{\sum x^2}{n} = \frac{7.70}{21} = .37$$

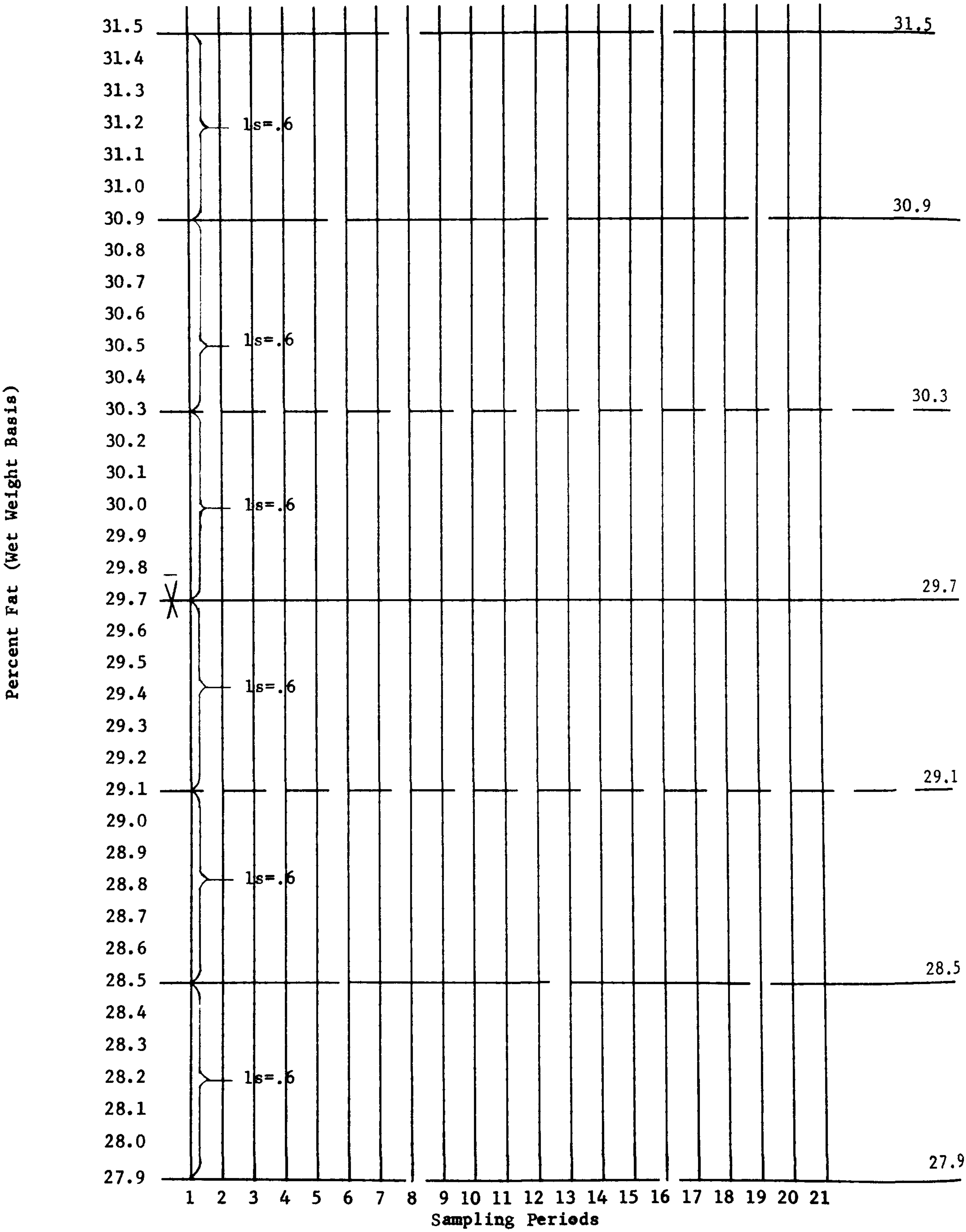
$$s^2 = \frac{\sum X^2}{n} - \bar{X}^2$$

$$= \frac{18{,}531.59}{21} - (29.7)^2$$

$$= 882.46 - 882.09 = .37$$

$$s = \sqrt{s^2} = \sqrt{.37} = .6$$

The fat percentages of this formulation has a mean of 29.7 and a standard deviation of 0.6 thus the following quality control chart can be constructed.



In future production, if the formulation is not altered the mean percent fat will be 29.7. Sixty-eight percent of the time the fat observation samples will be between 29.1 and 30.3%. Ninety-five percent of the time the observations will be between 28.5 and 30.9%. Ninety-nine percent of the time the observations will be between 27.9 and 31.5%.

If on repeated sampling a greater percentage of the observations than expected begin falling outside of these limits then the formulation has been altered and should be readjusted.

If this is not the desired range in fat percentage then a new formulation must be constructed and the calculations repeated. It should be noted that if a processor can reduce the variability (standard deviation) then the mean may be increased with the same percentage of observations falling under any or above any desired level. For example, if this meat processor wanted only 2.5% $\left[\frac{100-95}{2}\right]$ or actually 2.15% by previous normal curve of his product samples to exceed 30% fat content then he would have to reduce the mean fat value to 30 -2(.6) = 28.8 while maintaining the same quantity of variation (standard deviation of 0.6). This would necessitate a new quality control chart with; $\bar{X} = 28.8$ and $s = 0.6$. The same procedure could also be used with a minimum desired value.

Maximum Desired Value		Multiplier number can be estimated from normal curve or found precisely in a table	Calculated from previous production data

$$\text{Production } \bar{X} = \begin{bmatrix}\text{Maximum}\\ \text{desired}\\ \text{value}\end{bmatrix} - \begin{bmatrix}\text{Number of standard}\\ \text{deviations to give}\\ \text{desired \% above}\end{bmatrix} \bullet \begin{bmatrix}\text{Standard deviation}\end{bmatrix}$$

Minimum Desired Value

$$\text{Production } \bar{X} = \begin{bmatrix}\text{Minimum}\\ \text{desired}\\ \text{value}\end{bmatrix} + \begin{bmatrix}\text{Number of standard}\\ \text{deviations to give}\\ \text{desired \% below}\end{bmatrix} \bullet \begin{bmatrix}\text{Standard deviation}\end{bmatrix}$$

After the maximum (or minimum) desired value has been set and the percentage of individual samples that are permitted to exceed (or fall under) this value has been determined then the only thing that can change the production mean value ($\overline{X}$) is a change in the variability. For example, if the production variability could be changed in the previous frankfurter operation then the following production means could be used with the same percentage (2-1/2%) of observations exceeding the 30% fat level.

Production $\overline{X}$	Variability	Standard Deviation
28.8	No change	.6
29.4	Decreased variability	.3
28.0	Increased variability	1.0

A decrease in variability allows the moving of the production mean ($\overline{X}$) closer to the stated limits (30%).

QUALITY CONTROL CHARTS

There are several types of quality control charts that can be useful in the meat area. These are summarized in the next two tables and then discussed in more detail later in this chapter. The first job in setting up a quality control chart is deciding on the type of chart that can be used with the data that will be collected.

Type of Variable	Number of Samples	Type of Chart	Type of Data Plotted
Continuous	More than one	$\bar{X}$ - s	a) Individual values b) Mean of samples c) Standard deviation values
Continuous	More than one	$\bar{X}$ - R	a) Individual values b) Mean of samples c) Range of samples
Continuous	one	M $\bar{R}$	Individual values
Discrete	More than one (sample size varies from one sample to the next)	p - Chart	Fractional defective ; Proportion of defectives
Discrete	More than one (sample size the same for all samples)	P_n - Chart	Number of defections in each sample
Discrete	One sample (can have a large number of defects per sample)	c - Chart	Mean number of defects per sample

Types of Quality Control Charts

Type of Variable	Number of Samples Taken at a Time	Type of Chart	Sections of Chart and Type of Data Plotted: 99% Level	Sections of Chart and Type of Data Plotted: 95 % Level
Continuous	More than one	$\overline{X}$-s	a) Individual values $UCL(99)_X = \overline{X} + 3s$ or $\overline{X} + I_{1(99)}\ \overline{s} =$ $\overline{X} =$ $LCL(99)_X = \overline{X} - 3s$ or $\overline{X} - I_{1(99)}\ \overline{s} =$	$UCL(95)_X = \overline{X} + 2s$ or $\overline{X} + I_{1(95)}\overline{s} =$ $\overline{X} =$ $LCL(95)_X = \overline{X} - 2s$ or $\overline{X} + I_{1(95)}\overline{s} =$
			b) Mean of samples $UCL(99)_{\overline{X}} = \overline{\overline{X}} + 3s_{\overline{X}}$ or $\overline{\overline{X}}+A_{1(99)}\ \overline{s} =$ $\overline{\overline{X}} =$ $LCL(99)_{\overline{X}} = \overline{\overline{X}} - 3s_{\overline{X}}$ or $\overline{\overline{X}}-A_{1(99)}\ \overline{s} =$	$UCL(95)_{\overline{X}} = \overline{\overline{X}} + 2s_{\overline{X}}$ or $\overline{\overline{X}}+A_{1(95)}\overline{s} =$ $\overline{\overline{X}} =$ $LCL(95)_{\overline{X}} = \overline{\overline{X}} - 2s_{\overline{X}}$ or $\overline{\overline{X}}-A_{1(95)}\overline{s} =$
			c) Standard deviation values $UCL(99)_{s_i} = \overline{X}_s + 3_{s_s}$ or $B_{4(99)} \cdot \overline{s} =$ $\overline{X}_{s_i}$ or $\overline{s} =$ $LCL(99)_{s_i} = \overline{X}_s - 3_{s_s}$ or $B_{3(99)} \cdot \overline{s} =$	$UCL(95)_{s_i} = \overline{X}_s + 2_{s_s}$ or $B_{4(95)} \cdot \overline{s} =$ $\overline{X}_{s_i}$ or $\overline{s} =$ $LCL(95)_{s_i} = \overline{X}_s - 2_{s_s}$ or $B_{3(95)} \cdot \overline{s} =$

Types of Quality Control Charts

Type of Variable	Number of Samples Taken at a Time	Type of Chart	Sections of Chart and Type of Data Plotted: 99% Level	Sections of Chart and Type of Data Plotted: 95% Level
Continuous	More than one	$\bar{X}$-R	a) Individual values	
			$UCL(99)_X = \bar{\bar{X}} + I_{2(99)} (\bar{R}) =$ $\bar{X} =$ $LCL(99)_X = \bar{\bar{X}} - I_{2(99)} (\bar{R}) =$	$UCL(95)_X = \bar{\bar{X}} + I_{2(95)} \bar{R} =$ $\bar{X} =$ $LCL(95)_X = \bar{\bar{X}} - I_{2(95)} \bar{R} =$
			b) Mean of samples	
			$UCL(99)_{\bar{X}} = \bar{\bar{X}} + A_{2(99)} (\bar{R}) =$ $\bar{\bar{X}} =$ $LCL(99)_{\bar{X}} = \bar{\bar{X}} - A_{2(99)} (\bar{R}) =$	$UCL(95)_{\bar{X}} = \bar{\bar{X}} + A_{2(95)} (\bar{R}) =$ $\bar{\bar{X}} =$ $LCL(95)_{\bar{X}} = \bar{\bar{X}} - A_{2(95)} \bar{R} =$
			c) Range of samples	
			$UCL(99)_R = D_{4(99)} (\bar{R}) =$ $\bar{R} =$ $LCL(99)_R = D_{3(99)} (\bar{R}) =$	$UCL(95)_R = D_{4(95)} \bar{R} =$ $\bar{R} =$ $LCL(95)_R = D_{3(95)} \bar{R} =$

Types of Quality Control Charts

Type of Variable	Number of Samples Taken at a Time	Type of Chart	Sections of Chart and Type of Data Plotted: 99% Level	Sections of Chart and Type of Data Plotted: 95% Level
Continuous	One	$\overline{MR}$	Individual values $UCL(99)_X = \bar{X} + 2.66\ \overline{MR} =$ $\bar{X} =$ $LCL(99)_X = \bar{X} - 2.66\ \overline{MR} =$	$UCL(95)_{\bar{X}} = \bar{X} + 1.77\ \overline{MR} =$ $\bar{X} =$ $LCL(95)_{\bar{X}} = \bar{X} - 1.77\ \overline{MR} =$
Discrete	More than one (Sample size variable from one sample to the next)	P-Chart	Fractional defective $UCL(99)_p = \bar{p} + 3\sqrt{\frac{\bar{p}\ (1-\bar{p})}{n}} =$ $\bar{p} =$ $LCL(99)_p = \bar{p} - 3\sqrt{\frac{\bar{p}\ (1-\bar{p})}{n}} =$	$UCL(95)_p = \bar{p} + 2\sqrt{\frac{\bar{p}\ (1-\bar{p})}{n}} =$ $\bar{p} =$ $LCL(95)_p = \bar{p} - 2\sqrt{\frac{\bar{p}\ (1-\bar{p})}{n}} =$
Discrete	More than one (Sample size the same for all samples)	Pn Chart	Number of defections per sample $UCL(99)_{pn} = \overline{pn} + 3\sqrt{\overline{pn}(1-\bar{p})} =$ $\overline{pn} =$ $LCL(99)_{pn} = \overline{pn} - 3\sqrt{\overline{pn}(1-\bar{p})} =$	$UCL(95)_{pn} = \overline{pn} + 2\sqrt{\overline{pn}\ (1-\bar{p})} =$ $\overline{pn} =$ $LCL(95)_{pn} = \overline{pn} - 2\sqrt{\overline{pn}\ (1-\bar{p})} =$
Discrete	One sample (can have a large number of defects per sample)	c-Chart	Mean number of defects per sample $UCL(99)_c = \bar{c} + 3\sqrt{\bar{c}} =$ $\bar{c} =$ $LCL(99)_c = \bar{c} - 3\sqrt{\bar{c}} =$	$UCL(95) - \bar{c} + 2\sqrt{\bar{c}} =$ $\bar{c} =$ $LCL(95) = \bar{c} - 2\sqrt{\bar{c}} =$

$\bar{X}$-s: QUALITY CONTROL CHARTS

Since variability is so important, particularly from an economics standpoint, it is often useful to graph or plot a measure of variability (Example: standard deviation) with respect to time.

To calculate a periodic measure of variability more than one sample is normally taken at each sampling period. If the previous frankfurter fat percentage problem had been accomplished by taking 3 samples in each of 7 sampling periods (instead of one sample in each of 21 sampling periods) the results would look as follows:

Sampling Period

	A	B	C	D	E	F	G	Total
	29.4	29.8	29.7	28.7	30.0	29.9	30.3	207.8
	29.7	29.1	29.5	30.7	30.9	29.8	29.7	209.4
	30.5	28.5	29.2	29.5	29.7	28.9	30.2	206.5
Σ	89.6	87.4	88.4	88.9	90.6	88.6	90.2	623.70
$\bar{X}$	29.87	29.13	29.47	29.63	30.20	29.53	30.06	$\bar{\bar{X}} = 29.70$
ΣX^2	2676.70	2547.10	2604.98	2636.43	2736.90	2617.26	2712.22	
s^2	0.22	0.28	0.04	0.76	0.26	0.20	0.07	$\bar{X}_{s^2}$ or $\bar{s}^2 = 0.25$
s	0.46	0.53	0.20	0.82	0.51	0.45	0.26	$\bar{X}_s$ or $\bar{s} = 0.46$

$\bar{\bar{X}}$ = mean of the 7 individual mean values $= \dfrac{\bar{X}_A + \cdots + \bar{X}_G}{n \text{ of } \bar{X}\text{'s}}$

$$= \frac{29.87 + \cdots + 30.06}{7} = 29.70$$ = identical to $\bar{X}$ in value if there are no rounding errors

$\bar{X}_{s^2}$ or $\bar{s}^2$ = mean of the 7 individual variance values $= \dfrac{s_A^2 + \cdots + s_G^2}{n \text{ of } s^2\text{'s}} =$

$$= \frac{0.22 + \cdots + 0.07}{7} = 0.25$$

$\bar{X}_s$ or $\bar{s}$ = mean of the 7 individual standard deviation values $= \dfrac{s_A + \cdots + s_G}{7} =$

$$= \frac{0.46 + \cdots + 0.26}{7} = 0.46$$

In the preceding chapter an individual quality control chart (chart on which individual observations may be plotted) was constructed (99% level) in the following manner:

$$UCL(99)_X = \bar{X} + 3\ s = 29.7 + (3)\ (0.6) = 31.5$$

$$\bar{X} = 29.7$$

$$LCL(99)_X = \bar{X} - 3\ s = 29.7 - (3)\ (0.6) = 27.9$$

99% of original observations

The average standard deviation $\bar{X}_s$ or $\bar{s}$ calculated by determining the mean of the sampling periods standard deviation has slightly less (when a small sample size is used) variation than the total sample standard deviation (s).

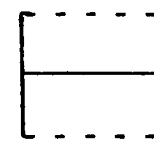
As the individual value standard deviations (s = 0.6; sample size 21) is compared to the mean standard deviation ($\bar{s}$ = 0.46; when sampled in groups of 3) it is apparent that as sample size decreases (21 to 3) that the standard deviation also decreases slightly. The magnitude of the group standard deviations depends on the sample groups, and in this example range from a minimum value of 0.20 to a maximum value of 0.82. When the mean of these values ($\bar{X}_s$ or $\bar{s}$ = 0.46) was calculated, the value obtained was 0.46 which generally should be slightly less than the standard deviation calculated from the original observation which was 0.6.

An alternate procedure for calculating the individual value chart from a mean standard deviation value ($\bar{s}$) compensates for this slight change (particularly with small sample sizes) in standard deviation and the general formula along with the frankfurter examples is as follows:

$$UCL(99)_X = \bar{X} + I_{1(99)}\,\bar{s} = 29.7 + (4.15)\ (0.46) = 29.7 + 1.9 = 31.6$$

$$\bar{X} = 29.7$$

$$LCL(99)_X = \bar{X} - I_{1(99)}\,\bar{s} = 29.7 - (4.15)\ (0.46) = 29.7 - 1.9 = 27.8$$



99% of original observations

See following I_1 table for these values.

The 95 and 99% levels on the frankfurter problem is now as follows:

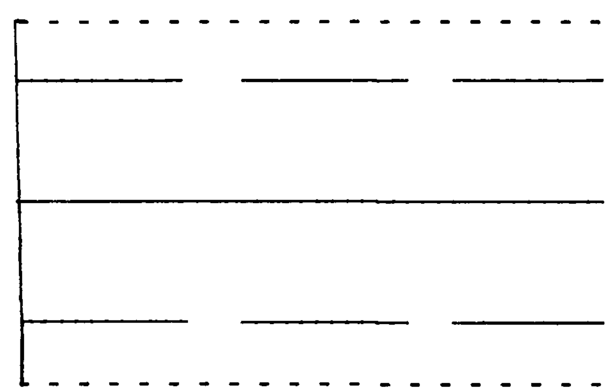
$UCL_{X(99)} = 29.7 + 4.15\ (0.46) = 31.61$

$UCL_{X(95)} = 29.7 + 2.77\ (0.46) = 30.97$

$\bar{X} = 29.7$

$LCL_{X(95)} = 29.7 - 2.77\ (0.46) = 28.43$

$LCL_{X(99)} = 29.7 - 4.15\ (0.46) = 27.79$



The 95 and 99% (2s and 3s) I_1 values can be obtained from the following table:

Table for Calculation of the Upper and Lower Control Limit for Individual Observations (X) From Mean Standard Deviation Values ($\bar{s}$).

Number of Samples Per Period	I_1 Value: $I_1(95)$= 95% level	$I_1(99)$= 99% level	Example of table use (This is not part of the frankfurter-fat problem)
2	3.55	5.32	$\bar{X} = 14;\ \bar{s} = 1.2$
3	2.77	4.15	
4	2.51	3.76	Number of samples in each period = 6
5	2.38	3.57	
6	2.30	3.45	What is the UCL & LCL for the individual observations at the 99% level?
7	2.25	3.38	
8	2.21	3.32	
9	2.19	3.28	UCL_X or individual values $= \bar{X} + I_1 (\bar{s})$
10	2.17	3.25	$= 14 + (3.45)\cdot(1.2)$
11	2.15	3.23	
12	2.14	3.21	$= \boxed{18.14}$
13	2.13	3.19	Target value $= \bar{X} = \boxed{14}$
14	2.11	3.17	LCL_X or individual values $= \bar{X} - I_1 (\bar{s})$
15	2.11	3.16	$= 14 - (3.45)\cdot(1.2)$
20	2.08	3.12	
25	2.06	3.09	$= \boxed{9.86}$

The confident limits for individual values for the frankfurter fat problem calculated from the overall standard deviation (s) and calculated from the I_1 values agrees very closely.

Individual Value (X) Control Limits

	Calculated from I_1 values		Calculated from the standard deviation (s)	
	95% level	99% level	95% level	99% level
UCL	30.97	31.61	29.7+2(.6)=30.90	29.7+3(.6)=31.50
LCL	28.43	27.79	29.7-2(.6)=28.50	29.7-3(.6)=27.90

The following factors should also be noted.

1. If the means for each sampling period $\left[(\bar{X}_A = 29.87) \text{ through } (\bar{X}_G = 30.06)\right]$ were plotted on the previously constructed control chart the sampling period means would fall between the overall mean (29.7) and plus or minus one (individual value) standard deviation (29.1 to 30.3). This would indicate that means (e.g., $X_A = 29.87$) have less variability than the individual observations (X). Logically this seems reasonable since an extreme value (most variation) would be grouped with less extreme values (values closer to the group mean). The sampling period mean (e.g., $\bar{X}_A$) would be closer to the total mean ($\bar{X}$) than would an individual (X) extreme observation.

2. A standard deviation calculated on the means ($s_{\bar{X}}$, calculated using the group means as original values were previously used and is sometimes referred to as error of the mean) yields a smaller value than the standard deviation calculated using the individual observations due to the reasons listed in #1. In this example the $s_{\bar{X}} = \sqrt{\frac{\sum \bar{X}^2}{n \text{ of } \bar{X}\text{'s}} - \bar{\bar{X}}^2}$ is calculated (see following) to be 0.34 and this value as suspected is less than the individual value standard deviation (s = 0.6)

Sampling Periods

	$\bar{X}_i^2$
$\bar{X}_A$ = 29.87	892.2169
$\bar{X}_B$ = 29.13	848.5569
$\bar{X}_C$ = 29.47	868.4809
$\bar{X}_D$ = 29.63	877.9369
$\bar{X}_E$ = 30.20	912.0400
$\bar{X}_F$ = 29.53	872.0209
$\bar{X}_G$ = 30.06	903.6036
$\Sigma\ \bar{X}$ = 207.89	$\Sigma\bar{X}^2$ 6174.8561

$\bar{\bar{X}}$ = 29.69857 (rounded to 29.70 on previous table)

$$s_{\overline{X}} = \sqrt{\frac{\sum \overline{X}^2}{\text{n of } \overline{X}\text{'s}} - \overline{\overline{X}}^2}$$

$$= \sqrt{\frac{6174.8561}{7} - (29.69857)^2}$$

$$= \sqrt{882.12230 - 882.00506}$$

$$= \sqrt{.11724}$$

$$= .34$$

The relationship between the standard deviation calculated from the sampling means (error of the mean) and the one calculated from the original observations is approximately (exact variation will depend on the way samples are grouped) as follows:

$$s_{\overline{X}} = \frac{s}{\sqrt{\text{n per group}}}$$

$$s = s_{\overline{X}} \cdot \sqrt{\text{n per group}}$$

$$s = 0.34 \text{ X } \sqrt{3}$$

$$s = 0.34 \text{ X } 1.7$$

$$s = 0.58$$

s calculated from original value = 0.6

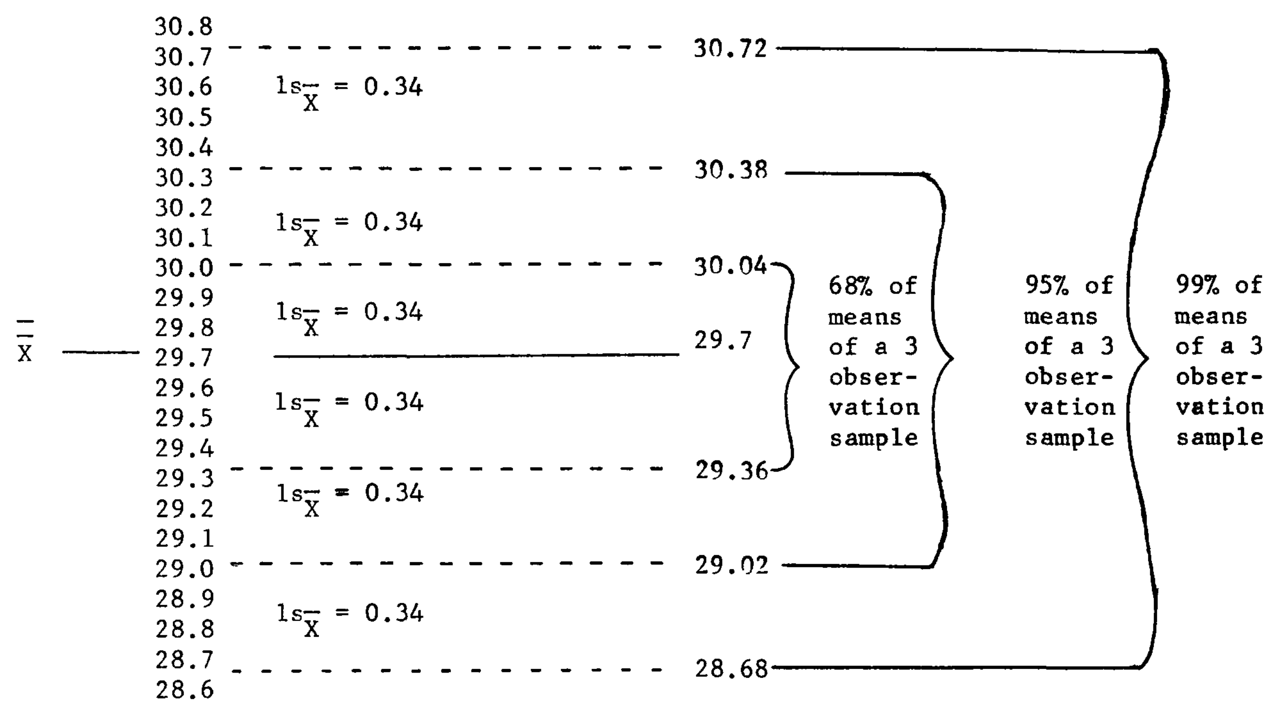
The relationship of these 2 values is also discussed in the chapter entitled "Significance", sub heading "Relationship of Variance between Population and Sample".

3. The standard deviation calculated on the sampling groups mean values (error of the mean) can be interpreted similarly to previously used standard deviations.

$\bar{\bar{X}} \pm 1\ s_{\bar{X}}$

29.7 ± 0.34

29.36 to 30.04 percent fat should contain 68% of the <u>means</u> of a 3 observation sample.



This type of quality control chart can be used to plot the mean value of a sampling period.

4. A quality control chart on which mean vlaues can be plotted can also be calculated using the mean of the group standard deviations $(\bar{s})$ and would be as follows:

$$UCL(99)_{\bar{X}} = \bar{X} + A_{1(99)}\ \bar{s} = 29.7 + (2.39)\ (0.46) = 30.80$$

$$\bar{X} = 29.7$$

$$LCL(99)_{\bar{X}} = \bar{X} - A_{1(99)}\ \bar{s} = 29.7 - (2.39)\ (0.46) = 28.60$$

99% of mean of a 3 observation sample

The 95 and 99% level on the frankfurter problem is now as follows:

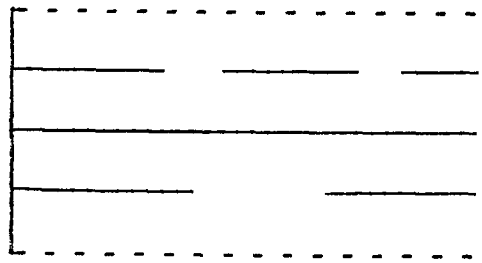
$UCL_{\bar{X}}$ (99) = 29.7 + 2.39 (0.46) = 30.80

$UCL_{\bar{X}}$ (95) = 29.7 + 1.59 (0.46) = 30.43

$\bar{X}$ = 29.7

$LCL_{\bar{X}}$ (95) = 29.7 - 1.59 (0.46) = 28.97

$LCL_{\bar{X}}$ (99) = 29.7 - 2.39 (0.46) = 28.60



The A_1 value for various sample sizes may be located in the following table:

Table for Calculating the Upper and Lower Control Limits for the Mean of Sampling Groups ($\bar{X}$)From Mean Standard Deviation Value ($\bar{s}$).

Number of Samples per Period	A_1 Value	
	A_1(95)=95% level	A_1(99)=99% level
2	2.51	3.76
3	1.59	2.39
4	1.25	1.88
5	1.07	1.60
6	0.94	1.41
7	0.85	1.28
8	0.79	1.18
9	0.73	1.09
10	0.69	1.03
11	0.65	0.97
12	0.61	0.92
13	0.59	0.88
14	0.57	0.85
15	0.55	0.82
20	0.47	0.70
25	0.41	0.62

Example of Table Use
(This is not part of the frankfurter-fat problem)

$\bar{\bar{X}} = 16$; $\bar{s} = 1.8$

Number of samples in each period = 7

What is the UCL & LCL for the mean of sampling group of size 7 at the 95% level?

$UCL_{\bar{X}} = \bar{\bar{X}} + A_1\ (\bar{s})$

$= 16 + (0.85)\ (1.8)$

$= \boxed{17.53}$

Target Value $\bar{\bar{X}} = \boxed{16}$

$LCL_{\bar{X}} = \bar{\bar{X}} - A_1\ (\bar{s})$

$= 16 - (0.85)\ (1.8)$

$= \boxed{14.47}$

The confidence limits for mean values for the frankfurter-fat problem calculated from the standard deviation of the sampling group means [error of the mean ($s_{\bar{X}}$)] and calculated from the A_1 values is in very close agreement.

Mean of Sample Values ($\bar{X}$) Control Limits

	Calculated from A_1 values		Calculated from error of the mean $s_{\bar{X}}$	
	95% level	99% level	95% level	99% level
UCL	30.43	30.80	29.7+2(.34)=30.38	29.7+3(.34)=30.72
LCL	28.97	28.60	29.7-2(.34)=29.02	29.7-3(.34)=28.68

5. The standard deviation values of the sampling periods are normally distributed. The variability of these values can be calculated by using the standard deviation formula with the group standard deviation values (e.g., s_A = .46) in place of the original observations. If this is done a standard deviation is calculated on the group standard deviations and will be symbolized as follows:

s_s and has a value of 0.19 in this example.

$$s_s = \sqrt{\frac{\sum (s)^2}{\#\text{ of } s} - \left(\bar{X}_s\right)^2}$$

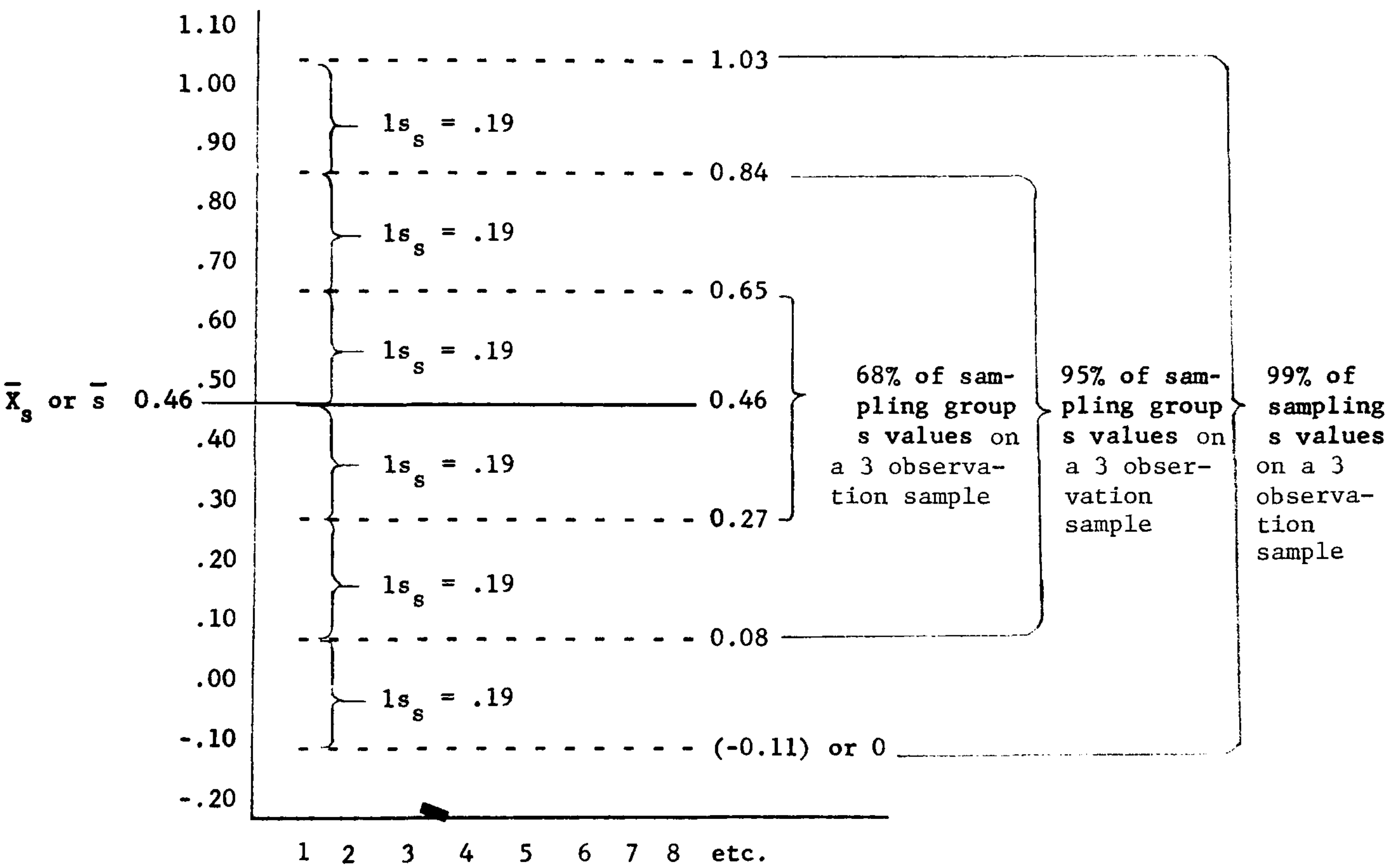
$$= \sqrt{\frac{1.7351}{7} - (.46)^2}$$

$$= \sqrt{.24787 - .2116}$$

$$= \sqrt{.03627}$$

$$= 0.19$$

This means that if groups of 3 samples are taken at a time and the standard deviation is calculated, 68% of these group standard deviations would be expected to be between 0.46 ± 0.19 (0.27 to 0.65). Since an increase in variability (group standard deviation) is going to cost the processor money and a decrease in variability is going to reduce cost it is quite helpful to plot the group standard deviations as follows:



Sampling Groups

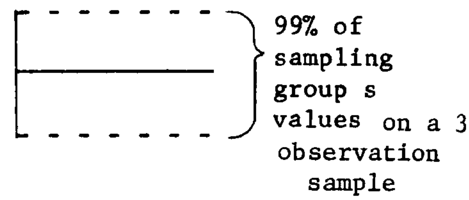
(3 samples per group)

6. A quality control chart on which group standard deviations can be plotted can also be calculated using the mean of the group standard deviations ($\bar{s}$) and would appear as follows:

$$UCL(99)_{s_i} = B_{4(99)} \cdot \bar{s} = 2.57 \cdot 0.46 = 1.18$$

$$\bar{s} = 0.46$$

$$LCL(99)_{s_i} = B_{3(99)} \cdot \bar{s} = 0 \cdot 0.46 = 0$$



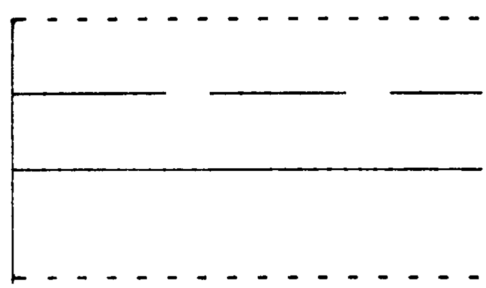
The 95 and 99% level on the frankfurter problem is now as follows:

$$UCL_{s(99)} = 2.57\ (0.46) = 1.18$$

$$UCL_{s(95)} = 2.04\ (0.46) = 0.94$$

$$\bar{s} = 0.46$$

$$\left.\begin{array}{l} LCL_{s(95)} \\ LCL_{s(99)} \end{array}\right\} = 0\ (0.46) = 0$$



The B_3 and B_4 values for various sample sizes may be located in the following table:

Table for Calculating the Upper and Lower Control Limits for Standard Deviations of Sample Groups (s_i) from the Mean Standard Deviation Value ($\bar{s}$)

Number of Samples Per Period	B_3 Value		B_4 Value	
	$B_{3(95)} =$ 95% level	$B_{3(99)} =$ 99% level	$B_{4(95)} =$ 95% level	$B_{4(99)} =$ 99% level
2	0	0	2.50	3.27
3	0	0	2.04	2.57
4	0.16	0	1.84	2.27
5	0.27	0	1.73	2.09
6	0.35	0.03	1.65	1.97
7	0.41	0.12	1.59	1.88
8	0.45	0.18	1.54	1.82
9	0.49	0.24	1.51	1.76
10	0.52	0.28	1.48	1.72
11	0.55	0.32	1.45	1.68
12	0.57	0.35	1.43	1.65
13	0.59	0.38	1.41	1.62
14	0.60	0.41	1.40	1.59
15	0.62	0.43	1.38	1.57
20	0.67	0.51	1.33	1.49
25	0.71	0.56	1.33	1.44

Example of table use (This is not a part of the frankfurter fat problem)

$\bar{s} = 2.2$

Number of samples in each period = 10

What is the UCL and LCL for standard deviations of the groups of size 10 at the 99% level?

$$UCL_{s_i} = B_4 \cdot \bar{s} = 1.72(2.2) = \boxed{3.78}$$

Target Value $= \bar{s} = \boxed{2.2}$

$$LCL_{s_i} = B_3 \cdot \bar{s} = 0.28(2.2) = \boxed{0.62}$$

The confidence limits for the standard deviations of the sample groups, for the frankfurter fat problem, calculated from the standard deviation of the group standard deviations (s_s) and calculated from the B_3 and B_4 values is in very close agreement.

Standard Deviation Values (s) Control Limits

	Calculated from B_3 and B_4 values		Calculated from standard deviation of group standard deviation (s_s)	
	95% level	99% level	95% level	99% level
UCL	.94	1.18	.46+2(.19)=.84	.46+3(.19)=1.03
LCL	0	0	.46-2(.19)=.08	.46-3(.19)= 0

Summary of $\bar{X}$-s Charts (99% level)

$\bar{X}$-s Chart	Calculations using individual value standard deviations, standard deviations of group means, and standard deviations of group standard deviations	Calculations using the mean of the standard deviations of the groups
	Individual Value: Chart	
Individual values plotted - - 3 per sampling period	$UCL(99)_X = \bar{X} + 3s = 31.5$ $\bar{X} = 29.7$ $LCL(99)_X = \bar{X} - 3s = 27.9$	$UCL(99)_X = \bar{X} + I_{1(99)}\,\bar{s} = 31.6$ $\bar{X} = 29.7$ $LCL(99)_X = \bar{X} - I_{1(99)}\,\bar{s} = 27.8$
	Mean of Sample Size 3: Chart	
Mean of 3 observations - - one plotted per each sampling period	$UCL(99)_{\bar{X}} = \bar{\bar{X}} + 3s_{\bar{X}} = 30.72$ $\bar{\bar{X}} = 29.7$ $LCL(99)_{\bar{X}} = \bar{\bar{X}} - 3s_{\bar{X}} = 28.68$	$UCL(99)_{\bar{X}} = \bar{\bar{X}} + A_{1(99)}\,\bar{s} = 30.8$ $\bar{\bar{X}} = 29.7$ $LCL(99)_{\bar{X}} = \bar{\bar{X}} - A_{1(99)}\,\bar{s} = 28.6$
	Standard Deviation: Chart (sample size 3)	
Standard deviation plotted - - one per sampling period	$UCL(99)_s = \bar{X}_{s_i} + 3s_s = 1.03$ $\bar{X}_s$ or $\bar{s} = 0.46$ $LCL(99)_s = \bar{X}_{s_i} - 3s_s = -0.11$	$UCL(99)_{s_i} = B_{4(99)}\,\bar{s} = 1.18$ $\bar{X}_s$ or $\bar{s} = 0.46$ $LCL(99)_{s_i} = B_{3(99)}\,\bar{s} = 0$

Practical Example of Control Chart

An example of how the information collected could be used in a frankfurter operation is as follows:

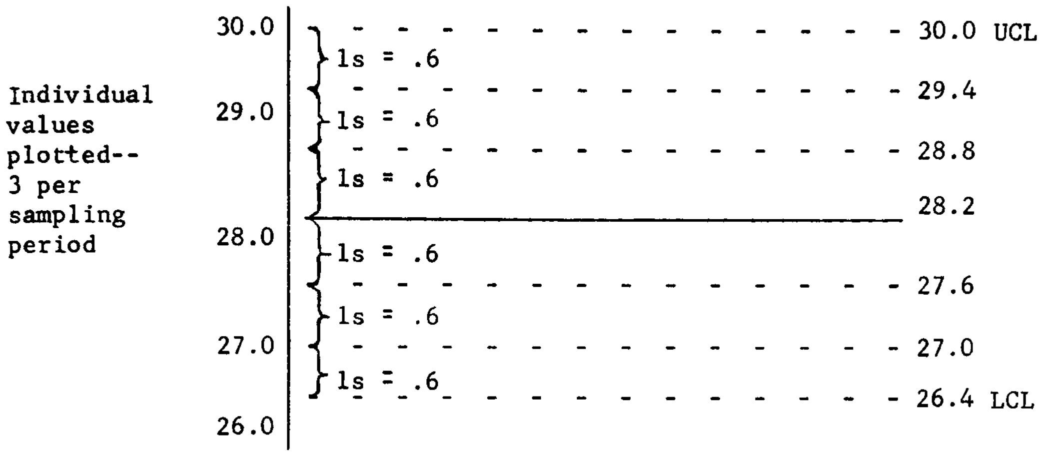
Company decisions:

1. 3 samples will be taken daily and analyzed for fat.
2. No more than 1/2 of 1% (3 standard deviations level) of the individual samples should exceed the 30% fat Federal inspection level.
3. Production will be adjusted when indicated by the quality control charts.

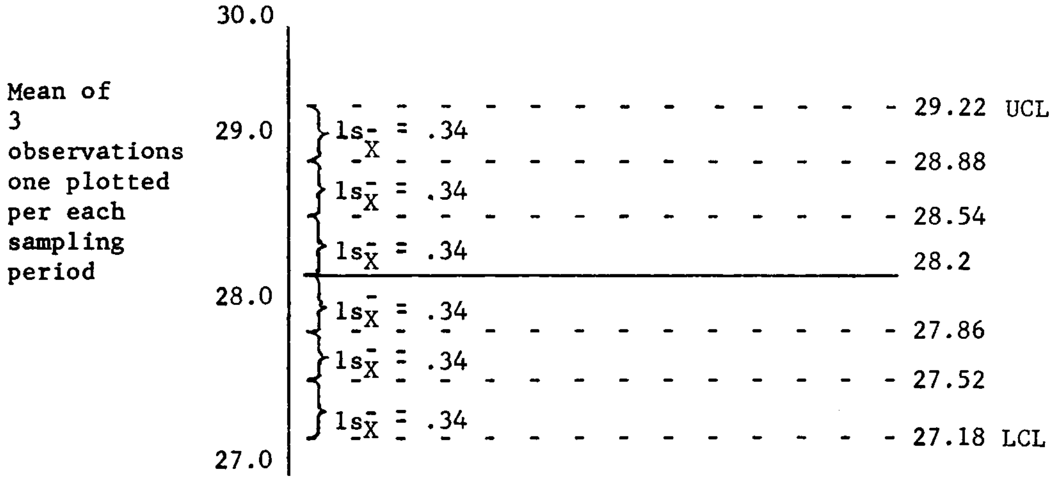
First chart that will be used based on information collected from the 21 previous samples.

$\overline{X}$-s Chart

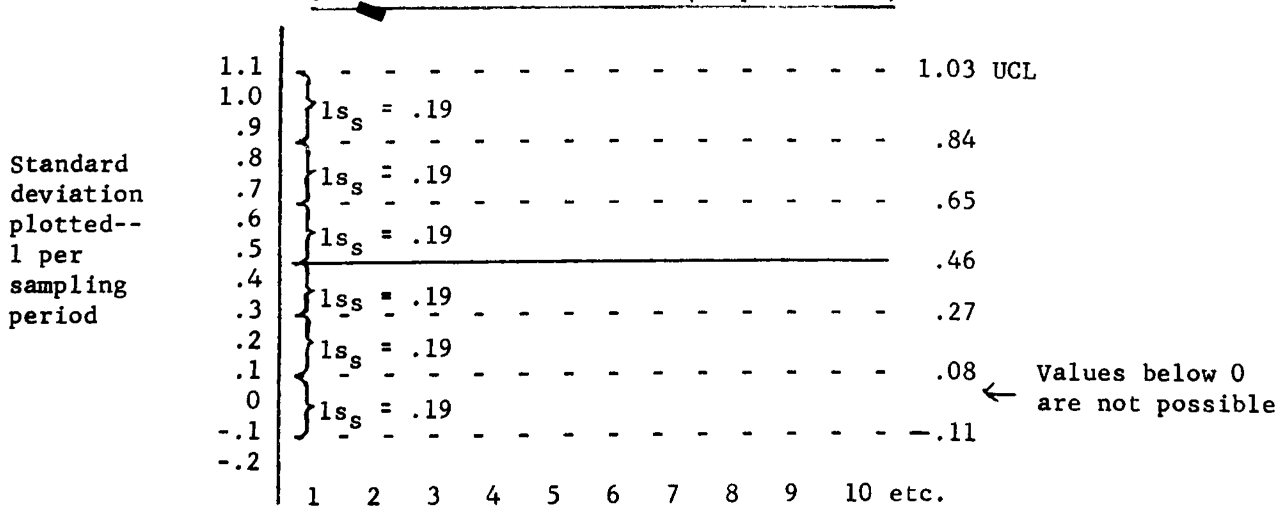
Individual Value Chart



Mean of Sample Size 3 Chart



Standard Deviation Chart (Sample size 3)



$\bar{X}$-R: QUALITY CONTROL CHART

Due to the quantity of math involved, the range is often used to construct a quality control chart instead of the standard deviation. This can be done by using tables or graphs just as the standard deviation was estimated from the range in a previous section of this text. If the "frankfurter - fat" samples were used the data could be arranged as follows:

Sampling Period	Samples per period			ΣX	$\bar{X}$	R
A	29.4	29.7	30.5	89.6	29.87	1.1
B	29.8	29.1	28.5	87.4	29.13	1.3
C	29.7	29.5	29.2	88.4	29.47	.5
D	28.7	30.7	29.5	88.9	29.63	2.0
E	30.0	30.9	29.7	90.6	30.20	1.2
F	29.9	29.8	28.9	88.6	29.53	1.0
G	30.3	29.7	30.2	90.2	30.06	.6
			Total	623.70	207.89	7.7
					$\bar{\bar{X}}$=29.70	$\bar{R}$=1.1

The 99% (3s) upper control limits (UCL) and lower control limits (LCL) for the individual values can be calculated as follows:

$$UCL(99)_{\text{Individual Values}} = \bar{\bar{X}} + I_{2(99)}\ (\bar{R})$$

$$LCL(99)_{\text{Individual Values}} = \bar{\bar{X}} - I_{2(99)}\ (\bar{R})$$

The I_2 values which are related to sample size can be obtained from the following table. With a sample size of 3 (in this example) the $I_{2(99)}$ value is 1.77.

Table for calculations of Upper & Lower Control Limits for <u>Individual Observations</u> (X) from the range data

Number of Samples per Period	I_2 value $\left[\frac{3}{d_2} \text{ or } \frac{2}{d_2}\right]$ (For explanation of I_2 values see appendix)	
	$I_{2(95)}$ = 95% level	$I_{2(99)}$ = 99% level
2	1.77	2.66
3	1.18	1.77
4	.97	1.46
5	.86	1.29
6	.79	1.18
7	.74	1.11
8	.70	1.05
9	.67	1.01
10	.65	0.98
11	.63	0.95
12	.61	0.92
13	.60	0.90
14	.59	0.88
15	.57	0.86
20	.53	0.80
25	.51	0.76

Example of table use (This is not part of the frankfurter-fat problem

$\bar{\bar{X}} = 10$; $\bar{R} = 2$

Number of samples in each period = 10

What is the UCL & LCL for the individual observations at the 99% level?

I_2 from table = 0.98

$UCL_X = \bar{\bar{X}} + I_2\ (\bar{R})$

$= 10 + .98(2)$

$= \boxed{11.96}$

Target Value = $\boxed{10}$

$LCL_X = \bar{\bar{X}} - I_2(\bar{R})$

$= 10 - .98(2)$

$= \boxed{8.04}$

Control limits (99%) for "frankfurter-fat problem":

$$UCL(99)_{\text{Individual Values}} = \bar{\bar{X}} + I_{2(99)}(\bar{R})$$
$$= 29.70 + 1.77(1.1)$$
$$= 29.70 + 1.95$$
$$= 31.65$$

$$LCL(99)_{\text{Individual Values}} = \bar{\bar{X}} - I_{2(99)}\bar{R}$$
$$= 29.70 - 1.77(1.1)$$
$$= 27.75$$

The 95 and 99% level on the frankfurter-fat problem is as follows:

$UCL_X(99) = 29.70 + 1.77(1.1) = 31.65$

$UCL_{X(95)} = 29.70 + 1.18(1.1) = 31.00$

$\bar{\bar{X}} = 29.70$

$LCL(95) = 29.70 - 1.18(1.1) = 28.40$

$LCL(99) = 29.70 - 1.77(1.1) = 27.75$

These control limits compare reasonably well with the ones calculated from the standard deviation(s) in the $\overline{X}$-s chart.

Individual Value (X) Control Limits

	Calculated from the range using I_2 values		Calculated from standard deviation(s)	
	95% Level	99% Level	95% Level	99% Level
UCL	31.00	31.65	29.7+2(.6)=30.90	29.7+3(.6)=31.50
LCL	28.40	27.75	29.7-2(.6)=28.50	29.7-3(.6)=27.90

The 99% (3s) Upper and Lower control limits for the mean values ($\bar{X}$) for each sampling group can also be calculated from the range in a similar manner and is illustrated as follows:

$$UCL(99)_{\bar{X}} \text{ values} = \bar{\bar{X}} + A_{2(99)} \ (\bar{R})$$

$$LCL(99)_{\bar{X}} \text{ values} = \bar{\bar{X}} - A_{2(99)} \ (\bar{R})$$

The $A_{2(99)}$ value varies with sample size and when multiplied by the average range ($\bar{R}$)is equal to 3 times the standard error of the mean (3 $s_{\bar{X}}$)

$$3 \ s_{\bar{X}} = A_{2(99)} \ \bar{R}$$

Therefore, the above formulas become:

$$CL_{\bar{X}} = \bar{\bar{X}} \pm A_2 \ \bar{R} \cong \bar{\bar{X}} \pm 3 \ s_{\bar{X}}$$

The A_2 value can be obtained from the following table. With a sample size of 3 (in this example) the A_2 value at the 99% level is 1.02.

Table for calculation of upper and lower control limits for means ($\bar{X}$) from the range data

Number of Samples Per Period	A_2 Value For explanation of A_2 see appendix $\left[\frac{3}{\sqrt{n}\,(d_2)}\right]$ or $\left[\frac{2}{\sqrt{n}\,(d_2)}\right]$ $A_{2(95)}$=95% level	$A_{2(99)}$=99% level	Example of table use (This is not part of the "frankfurter-fat" problem)
2	1.25	1.88	$\bar{\bar{X}}$ = 12; $\bar{R}$ = 1.5
3	.68	1.02	
4	.49	0.73	Number of samples per
5	.39	0.58	period = 5
6	.32	0.48	What is the UCL & LCL for
7	.28	0.42	the mean values ($\bar{X}$) at the
8	.25	0.37	95% level?
9	.23	0.34	A_2 from table = 0.39
10	.21	0.31	$UCL_{\bar{X}} = \bar{\bar{X}} + A_2(\bar{R})$
11	.19	0.29	$= 12 + .39(1.5)$
12	.18	0.27	$= \boxed{12.59}$
13	.17	0.25	Target value = $\bar{\bar{X}} = \boxed{12}$
14	.16	0.24	$LCL_{\bar{X}} = \bar{\bar{X}} - A_2(\bar{R})$
15	.15	0.22	$= 12 - .39(1.5)$
20	.12	0.18	$= \boxed{11.42}$
25	.10	0.15	

$$\begin{aligned} UCL(99)_{\bar{X}} \text{ Values} &= \bar{\bar{X}} + A_{2(99)}(\bar{R}) \\ &= 29.70 + 1.02(1.1) \\ &= 29.70 + 1.12 \\ &= 30.82 \end{aligned}$$

$$\begin{aligned} LCL(99)_{\bar{X}} \text{ Values} &= \bar{\bar{X}} - A_{2(99)}(\bar{R}) \\ &= 29.70 - 1.02(1.1) \\ &= 28.58 \end{aligned}$$

The 95 and 99% level on the frankfurter-fat problem is as follows:

$$UCL(99)_{\bar{X}} = 29.70 + 1.02(1.1) = 30.82$$
$$UCL(95)_{\bar{X}} = 29.70 + 0.68(1.1) = 30.45$$
$$\bar{\bar{X}} = 29.70$$
$$LCL(95)_{\bar{X}} = 29.70 - 0.68(1.1) = 28.95$$
$$LCL(99)_{\bar{X}} = 29.70 - 1.02(1.1) = 28.58$$

These control limits compare reasonably well with those calculated from the standard deviation(s) in the X-s chart.

Mean Sample Values (X) Control Limits

	Calculated from the range using A_2 values		Calculated from the error of the mean $s_{\bar{X}}$	
	95% Level	99% Level	95% Level	99% Level
$UCL_{\bar{X}}$	30.45	30.82	29.7+2(.34)=30.38	29.7+3(.34)=30.72
$LCL_{\bar{X}}$	28.95	28.58	29.7-2(.34)=29.02	29.7-3(.34)=28.68

The 99% (3s) upper and lower control limits for the range from each sampling group can also be calculated from the range data in a similar manner and is illustrated as follows:

$$UCL(99)_{R\ value} = D_{4(99)}(\bar{R})$$

$$LCL(99)_{R\ value} = D_{3(99)}(\bar{R})$$

The D_4 and D_3 values can be obtained from the following table. With a sample size of 3 (in this example) the $D_{4(99)}$ is 2.57 and $D_{3(99)}$ value is 0.

Table for Calculation of Upper and Lower Control Limits for the Range Values (R) from Range Data

Number of Samples per Period	D_4 Value Upper $\left[1 + \frac{3s_w}{d_2} \text{ or } 1 + \frac{2s_w}{d_2}\right]$ For explanation of D_4 see appendix		D_3 Value Lower $\left[1 - \frac{3s_w}{d_2} \text{ or } 1 - \frac{2s_w}{d_2}\right]$ For explanation of D_3 see appendix	
	$D_{4(95)}$ = 95% Level	$D_{4(99)}$ = 99% Level	$D_{3(95)}$ = 95% Level	$D_{3(99)}$ = 99% Level
2	2.51	3.27	0	0
3	2.05	2.57	0	0
4	1.85	2.28	.15	0
5	1.74	2.11	.26	0
6	1.67	2.00	.33	0
7	1.62	1.92	.38	0.08
8	1.58	1.86	.42	0.14
9	1.54	1.82	.46	0.18
10	1.52	1.78	.48	0.22
11	1.50	1.74	.50	0.26
12	1.48	1.72	.52	0.28
13	1.46	1.69	.54	0.31
14	1.45	1.67	.55	0.33
15	1.43	1.65	.57	0.35
20	1.39	1.59	.61	0.42
25	1.36	1.54	.64	0.46

Example of Table Use (This is not part of the "frankfurter-fat" problem)

$\bar{R} = 3.0$

Number of samples per period = 8

What is the UCL & LCL for the Range (R) at the 99% level?

D_4 from table = 1.86

$UCL_R = D_4(\bar{R})$

$= 1.86(3.0)$

$= \boxed{5.58}$

Target Value = $\bar{R}$

$= \boxed{3}$

D_3 from table = 0.14

$LCL_R = D_3(\bar{R})$

$= 0.14(3.0)$

$= \boxed{0.42}$

$$UCL(99)_R \text{ Value} = D_{4(99)}(\overline{R}) = 2.57(1.1) = 2.83$$

$$LCL(99)_R \text{ Value} = D_{3(99)}(\overline{R}) = 0(1.1) = 0$$

The 95 and 99% level on the frankfurter problem is as follows:

$$UCL(99)_R = 2.57(1.1) = 2.83$$

$$UCL(95)_R = 2.05(1.1) = 2.26$$

$$\overline{R} = 1.1$$

$$\left.\begin{array}{l} LCL(95)_R = \\ LCL(99)_R = \end{array}\right\} 0(1.1) = 0$$

The total $\overline{X}$-R control chart at the 99% level can now be illustrated.

$\overline{X}$-R Chart (99% level) for Frankfurter-Fat Example

			Individual Value--Chart
Individual Value Plotted-- 3 Per Sampling Period	UCL_X	31.65	- - - - - - - -
	$\overline{X}$	29.70	————————
	LCL_X	27.75	- - - - - - - -
			Mean of Sample Size 3--Chart
Mean of 3 Observations-- One Plotted Per Each Sampling Period	$UCL_{\overline{X}}$	30.82	- - - - - - - -
	$\overline{X}$	29.70	————————
	$LCL_{\overline{X}}$	28.58	- - - - - - - -
			Range Chart (Sample Size 3)
Range Plotted-- 1 Per Sampling Period	UCL_R	2.83	- - - - - - - -
	$\overline{R}$	1.1	————————
	LCL_R	0	- - - - - - - -

MOVING RANGE ($\overline{MR}$) CHARTS

Sometimes it is impractical or impossible to secure more than one sample at a time. In this case a moving range chart is used for control and control limits are based on a moving range of the individual measurements. In the "frankfurter-fat" example if 21 samples were taken one at a time the results would look as follows:

X	MR difference between adjacent samples
29.4	
	.3
29.7	
	.8
30.5	
	.7
29.8	
	.7
29.1	
	.6
28.5	
	1.2
29.7	
	.2
29.5	
	.3
29.2	
	.5
28.7	
	2.0
30.7	
	1.2
29.5	
	.5
30.0	
	.9
30.9	
	1.2
29.7	
	.2
29.9	
	.1
29.8	
	.9
28.9	
	1.4
30.3	
	.6
29.7	
	.5
30.2	

$\sum X = 632.7$

$n = 21$

$\overline{X} = \frac{\sum X}{n} = \frac{623.7}{21}$

$= 29.70$

MR 14.8

$\overline{MR} = \frac{RM}{n-1}$

$= \frac{14.8}{21-1}$

$= 0.74$

99% (3s) control limits on individual observations (X) using MR

Control limits$(99)_X = \overline{X} \pm 3s$

s is estimated from the range by

$$s = \frac{\overline{MR}}{d_2}$$

Therefore: Control limits $(99)_X$ =

$$\overline{X} \pm 3\,\frac{\overline{MR}}{d_2} = \overline{X} \pm \frac{3}{d_2}\quad(\overline{MR})$$

In the MR the n of observation (adjacent samples) is 2

Using the graph in the preceding section d_2 (for sample size of 2)=1.128

Control limits$(99)_X = \overline{X} \pm \frac{3}{1.128}\,\overline{MR}$ for 99% level

$\boxed{\text{Control Limits}(99)_X = \overline{X} \pm 2.66\ \overline{MR} \text{ for 99\% level}}$

Control limits$(95)_X = \overline{X} \pm \frac{2}{1.128}\cdot\overline{MR}$ for 95% level

$\boxed{\text{Control Limits}(95)_X = \overline{X} \pm 1.77\cdot\overline{MR} \text{ for 95\% level}}$

In the frankfurter-fat problem (99% level):

$$\begin{aligned} UCL(99)_X &= \overline{X} + 2.66\ \overline{MR} \\ &= 29.70 + 2.66\ (.74) \\ &= 29.70 + 1.97 \\ &= 31.67 \end{aligned}$$

$$\begin{aligned} LCL(99)_X &= \overline{X} - 2.66\ \overline{MR} \\ &= 29.70 - 2.66\ (.74) \\ &= 27.73 \end{aligned}$$

The control chart for this example would look as follows:

$\overline{MR}$ Chart - Frankfurter-fat (one sample taken at a time)

$$UCL(99)_X = 29.70 + 2.66(0.74) = 31.67$$

$$UCL(95)_X = 29.70 + 1.77(0.74) = 31.01$$

$$\bar{X} = 29.70$$

$$LCL(95)_X = 29.70 - 1.77(0.74) = 28.39$$

$$LCL(99)_X = 29.70 - 2.66(0.74) = 27.73$$

Sampling Periods

Again, these control limits compare reasonably well with the ones calculated from the standard deviation.

Individual Value (X) Control Limits

	Calculated From $\overline{MR}$		Calculated from Standard Deviattion(s)	
	95% Level	99% Level	95% Level	99% Level
UCL_X	31.01	31.67	29.7+2(.6)=30.90	29.7+3(.6)=31.50
LCL_X	28.39	27.73	29.7-2(.6)=28.50	29.7-3(.6)=27.90

A control chart made from the moving range only has one section for plotting individual values. The standard deviation of the moving range values could be calculated but in actual practice this is not often done.

p-CHARTS (Fraction Defective)

Charts of this nature are used when the results are counted (discrete variable) rather than measured (continuous variable).

$$\text{fraction defective} = \frac{\text{number of defective samples}}{\text{total number of samples examined}}$$

If 200-one pound samples of boneless beef were examined and 2 samples were found to contain a critical defect then the fraction defective would be as folllows:

$$p = \frac{h}{n}$$

$$p = \frac{2}{200} = .01 \text{ or } 1\%$$

Example of construction of a p-chart.

Sampling Period	Sample Size (No. of one pound samples n	Number with critical defects h	Fraction defective $p = \frac{h}{n}$
A	100	0	0
B	150	1	.0067
C	200	2	.01
D	150	0	0
E	100	2	.02
F	150	0	0
G	200	1	.005
H	100	1	.01
I	100	0	0
J	200	3	.015

$\sum n = 1450$ $\sum h = 10$

$$\bar{p} = \frac{\sum h}{\sum n}$$

$$= \frac{10}{1450} \doteq .0069$$

99% level

95% level

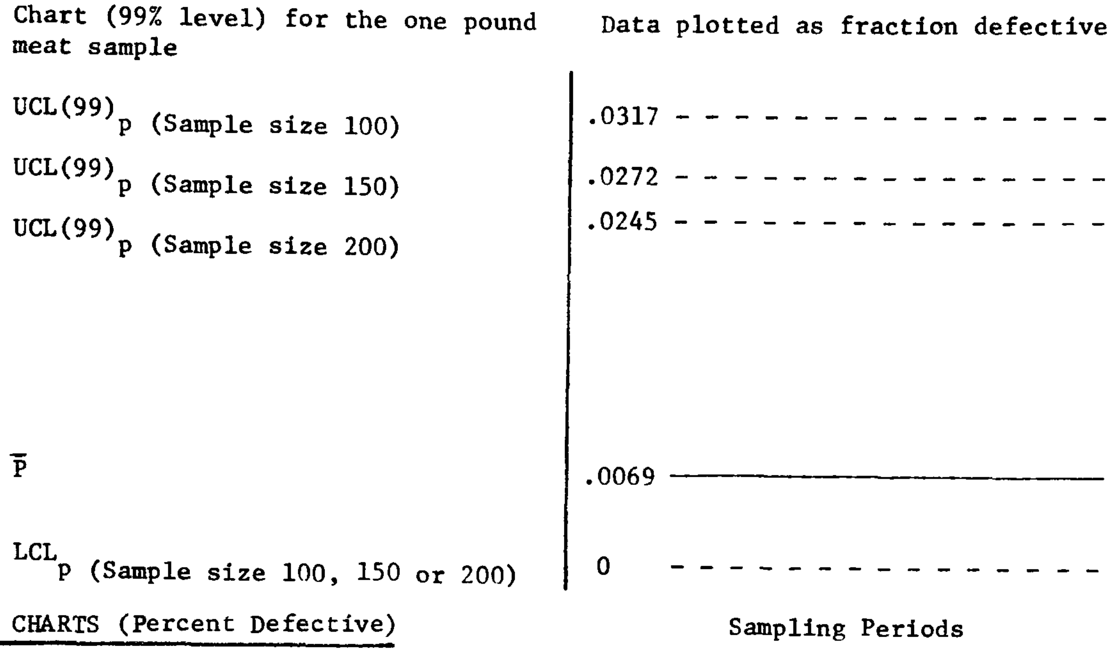
$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \quad \text{or} \quad \bar{p} + 2\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \quad \text{or} \quad \bar{p} - 2\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

Calculation of Upper & Lower Control Limits for Various Size Samples
(In this example 100, 150, & 200)

Sample Size	$UCL(99)_p$	$LCL(99)_p$
General Formula	$UCL(99)_p = \bar{p} + 3\sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}}$	$LCL(99)_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$
100	$UCL(99)_p = .0069 + 3\sqrt{\frac{(.0069)(1-.0069)}{100}}$ $= .0069 + 3\sqrt{\frac{(.0069)(.9931)}{100}}$ $= .0069 + 3\sqrt{\frac{.00685}{100}}$ $= .0069 + 3\sqrt{.000068}$ $= .0069 + 3\,(.0083)$ $= .0069 + .0248$ $= .0317$ $\boxed{UCL(95)_p = .0235}$	$LCL(99)_p = .0069 - 3\sqrt{\frac{(.0069)(1-.0069)}{100}}$ $= .0069 - .0248$ $= -.0179$ or 0 $\boxed{LCL(95)_p = -.0097 \text{ or } 0}$
150	$UCL(99)_p = .0069 + 3\sqrt{\frac{(.0069)(1-.0069)}{150}}$ $= .0069 + 3\sqrt{\frac{.00685}{150}}$ $= .0069 + 3\sqrt{.000045}$ $= .0069 + 3(.0068)$ $= .0069 + .0203$ $= .0272$ $\boxed{UCL(95)_p = .0205}$	$LCL(99)_p = .0069 - 3\sqrt{\frac{(.0069)(1-.0069)}{150}}$ $= .0069 - .0203$ $= -.0134$ or 0 $\boxed{LCL(95)_p = -.0067 \text{ or } 0}$
200	$UCL(99)_p = .0069 + 3\sqrt{\frac{(.0069)(1-.0069)}{200}}$ $= .0069 + 3\sqrt{\frac{.00685}{200}}$ $= .0069 + 3\sqrt{.000034}$ $= .0069 + 3(.0059)$ $= .0069 + .0176$ $= .0245$ $\boxed{UCL(95)_p = .0187}$	$LCL(99)_p = .0069 - 3\sqrt{\frac{(.0069)(1-.0069)}{200}}$ $= .0069 - .0176$ $= -.0107$ or 0 $\boxed{LCL(95)_p = -.0049 \text{ or } 0}$

P - Chart (99% level) for the one pound meat sample



pn - CHARTS (Percent Defective)

Percent defective charting can be simplified if the same size sample is always taken. If the sample size is constant then a pn chart may be used instead of a p chart.

Example of the Construction of a pn Chart

Sampling Period	Sample Size n	Number of Defects in Sample* h	Fraction Defective* $p = \frac{h}{n}$	Number of Defectives $pn = (\frac{h}{n})n = h$
A	50	0	0	0
B	50	2	.04	2
C	50	1	.02	1
D	50	0	0	0
E	50	0	0	0
F	50	3	.06	3
G	50	1	.02	1
H	50	1	.02	1
I	50	0	0	0
J	50	0	0	0
$\Sigma n = 500$		$\Sigma h = 8$		$\Sigma pn = 8$

* Column not necessary for calculation

$\bar{p} = \frac{\Sigma pn}{\Sigma n} = \frac{8}{500} = .016$ $\qquad$ $\bar{pn} = \frac{\Sigma pn}{\text{\# of sample periods}} = \frac{8}{10} = .8$

$\boxed{\text{99\% Level}}$		$\boxed{\text{95\% Level}}$
$UCL_{pn} = \overline{pn} + 3\sqrt{\overline{pn}(1-\overline{p})}$	or	$\overline{pn} + 2\sqrt{\overline{pn}(1-\overline{p})}$
$= .8 + 3\sqrt{.8(1-.016)}$		$.8 + 2\sqrt{.8(1-.016)}$
$= .8 + 3\sqrt{.8(.984)}$		$.8 + 2\sqrt{.8(.984)}$
$= .8 + 3\sqrt{.7872}$		$.8 + 2\sqrt{.7872}$
$= .8 + 3\ (.8872)$		$.8 + 2\ (.8872)$
$= .8 + 2.66$		$.8 + 1.77$
$= \boxed{3.46}$		$\boxed{2.57}$
Target Value $\overline{pn} = \boxed{0.8}$	or	$\boxed{0.8}$
$LCL_{pn} = \overline{pn} - 3\sqrt{\overline{pn}(1-\overline{p})}$	or	$\overline{pn} - 2\sqrt{\overline{pn}(1-\overline{p})}$
$= .8 - 3\sqrt{.8(1-.016)}$		$.8 - 2\sqrt{.8(1-.016)}$
$= .8 - 2.66$		$.8 - 1.77$
$= \boxed{-1.86 \text{ or } 0}$		$\boxed{-0.97 \text{ or } 0}$

pn Chart for the Previous One Pound Samples (Sample size of 50)

	Number of defective items per observations plotted at 99% level		Number of defective items per observations plotted at 95% level
$UCL(99)_{pn}$	3.46 - - - - - - -	$UCL(95)_{pn}$	2.57 - - - - - - - -
$\overline{pn}$	.8 ————————	$\overline{pn}$	.8 ————————
$LCL(99)_{pn}$	0 - - - - - - - -	$LCL(95)_{pn}$	0 - - - - - - - -
	Sampling Period		Sampling Period

<u>c-Chart</u>

c - Charts are used when a large number of defects are possible in any one item, for example the number of hairs found on a beef carcass. If a frequency distribution is plotted from this type of data a normal distribution will not be obtained but the curve will be positive skewed and is called a Poisson distribution.

A control c-chart would be calculated as follows:

Sampling Period	Sample Size	Number of Defects per Sample c
A	1	3
B	1	2
C	1	7
D	1	1
E	1	0
F	1	2
G	1	1
H	1	3
I	1	10
J	1	1
	$\Sigma n = 10$	$\Sigma c = 30$

$$\bar{c} = \frac{\Sigma c}{\Sigma n} = \frac{30}{10} = 3$$

	99% Level		95% Level
UCL_c	$= \bar{c} + 3\sqrt{\bar{c}}$	or	$\bar{c} + 2\sqrt{\bar{c}}$
	$= 3 + 3\sqrt{3}$		$3 + 2\sqrt{3}$
	$= 3 + 3(1.73)$		$3 + 2(1.73)$
	$= 3 + 5.20$		$3 + 3.46$
	$= \boxed{8.20}$		$\boxed{6.46}$
Target Value $= \bar{c}$	$= \boxed{3}$		$\boxed{3}$
LCL_c	$= \bar{c} - 3\sqrt{\bar{c}}$	or	$\bar{c} - 2\sqrt{\bar{c}}$
	$= 3 - 3\sqrt{3}$		$3 - 2\sqrt{3}$
	$= 3 - 5.20$		$3 - 3.46$
	$= -2.2$ or 0		-0.46 or 0

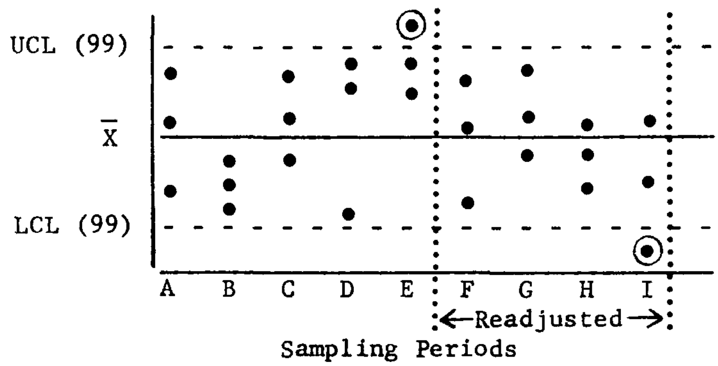
c-Chart for previous example (defects counted for one sample)

	Number of defects plotted per sample at 99% level		Number of defects plotted per sample at 95% level
$UCL(99)_c$	8.2 - - - - - - - - - - - -	$UCL(95)_c$	6.46 - - - - - - - - - - - -
$\bar{c}$	3 ————————————	$\bar{c}$	3 ————————————
$LCL(99)_c$	0 - - - - - - - - - - - -	$LCL(95)_c$	0 - - - - - - - - - - - -
	Sampling Period		Sampling Period

DETECTING & DIAGNOSING FROM A QUALITY CONTROL CHART

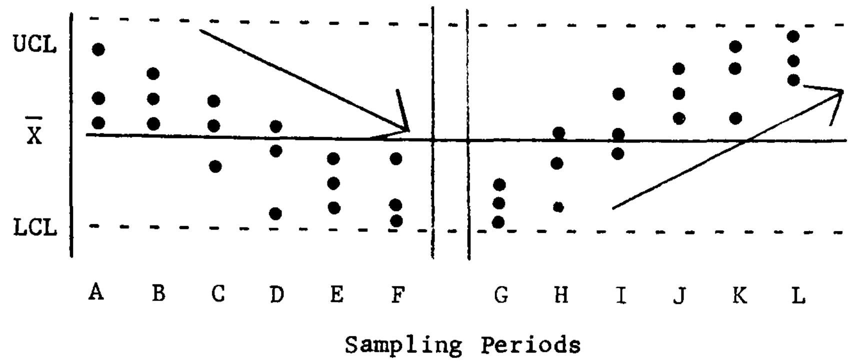
Once a quality control chart has been constructed then it should be observed very closely for indications of when the process is out of control or for additional indicators that can be obtained about the population. A few of these indicators are outlined below.

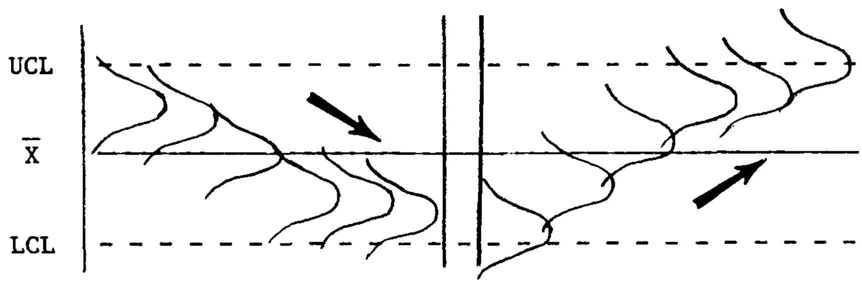
1. An observation exceeding the upper or lower control limit.



   Since less than 0.5% (one out of 200 samples) of our samples should exceed the 3σ UCL (99) (upper control limits) the process is probably out of control at Sampling Period E and should be readjusted by reducing the Mean Value at that Point. At Sampling Period I the process seems to be out of control again and the mean value should be adjusted upward.

2. If a trend can be observed without a sample exceeding the upper control limits (UCL) or the lower control limits (LCL) adjustments may be made to bring the operation back into control prior to it actually exceeding these limits. Generally 6 consecutive sampling periods in one direction are considered a minimum to establish a trend. Trends may be noted in the following examples.



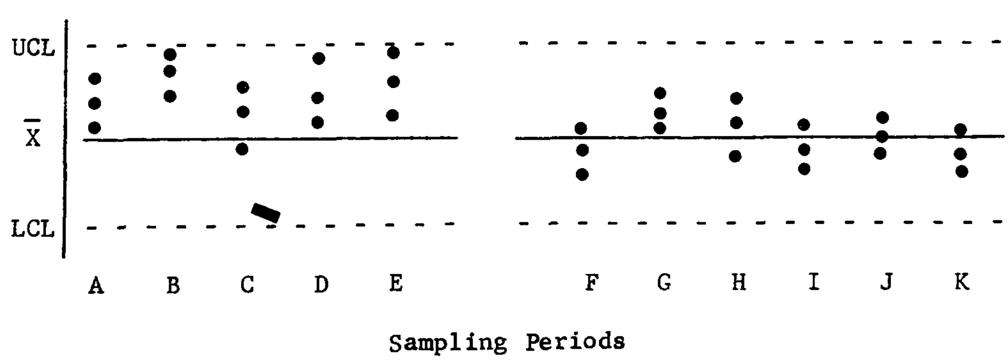


Level is declining. Should be adjusted upward.

Level is increasing. Should be adjusted downward.

When a trend is noted it is an indication that something is affecting the population mean at a fairly slow constant rate.

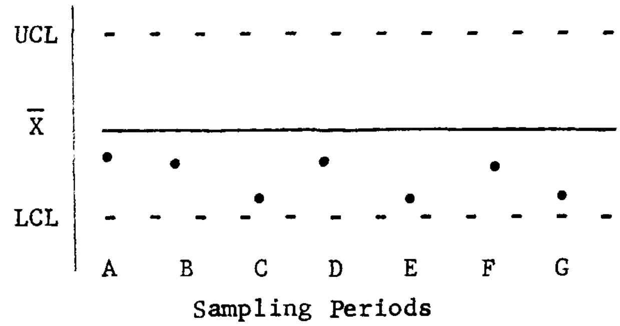
3. Often variability information can also be observed on these individual value control charts.



Level is above mean. Variability also seems to be reduced and can be checked on group standard deviation or range chart. If this is verified then the desired production mean can be increased or reduced and new control limits calculated

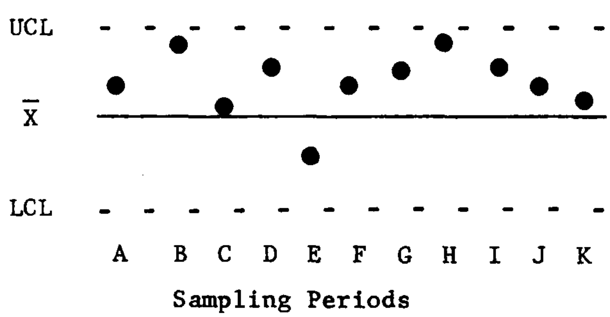
Level is near mean value. Variability also seems to be reduced and can be checked on group standard deviation or range chart. If this is verified then a new control chart should be calculated and the mean level could be adjusted if desired and new control limits calculated

4. Consecutive values on the same side of the central value. The probability of one value being either above or below the center value is 1/2.



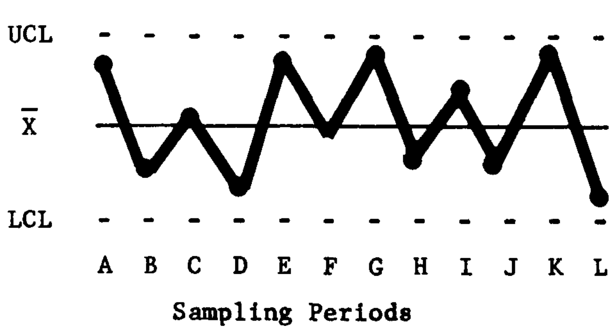
The probability of 7 consecutive values being on the same side of the center value is: (1/2)(1/2)(1/2)(1/2)(1/2)(1/2)(1/2) = 1/128. The mean of the population has shifted downward and should be readjusted even though no samples have exceeded the limits.

5. Ten out of eleven successive plots on the same side of the center line.



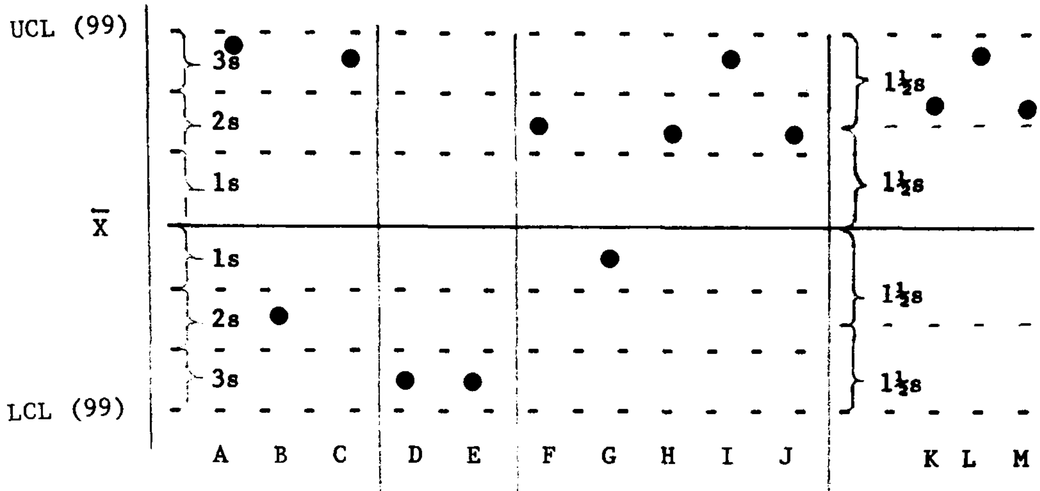
The probability of 10 out of 11 or 7 consecutive at the same point in time is $\left[(1/2)^{11}\right] 4 = \frac{1}{512}$.

6. Systematic variable - alternate high and low samples.



High and low samples in a regular sequence (saw tooth) indicate that there is a systematic variable causing this variation.

7. If the upper and lower control limits are calculated at the 3s level the graphs can then be divided into 6 portions as follows:

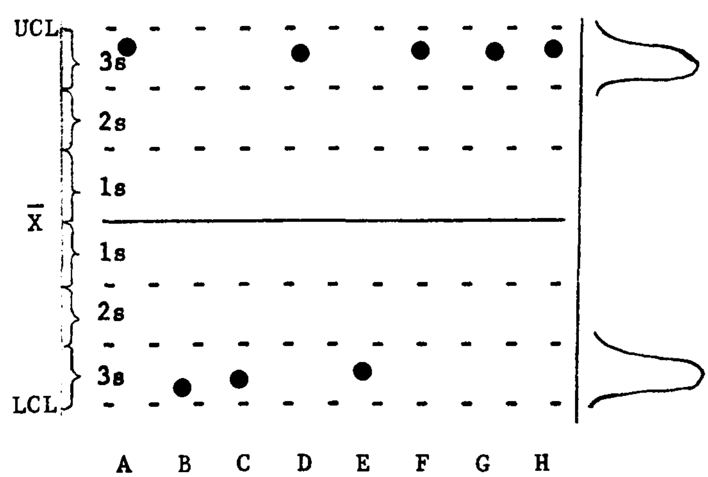


2 out of 3 in or beyond the 3s band [P of 2 out of 3 or 2 consecutive at some point in time = (1/20 x 1/20 x 19/20)2 = 19/4000]	2 consecutive in or beyond the 3s band [(P = 1/20 x 1/20 = 1/400]	4 out of 5 in or beyond the 2s band [P = $\{(1/3)^4 \cdot 2/3\}$ 5 = 10/243]	3 consecutive in or beyond the 1-1/2s line [P = $(.07)^3$ = .0003]

Each of these are indications that the process is out of control.

In any of these 4 cases the process should be adjusted even though no observation may be outside the control limits.

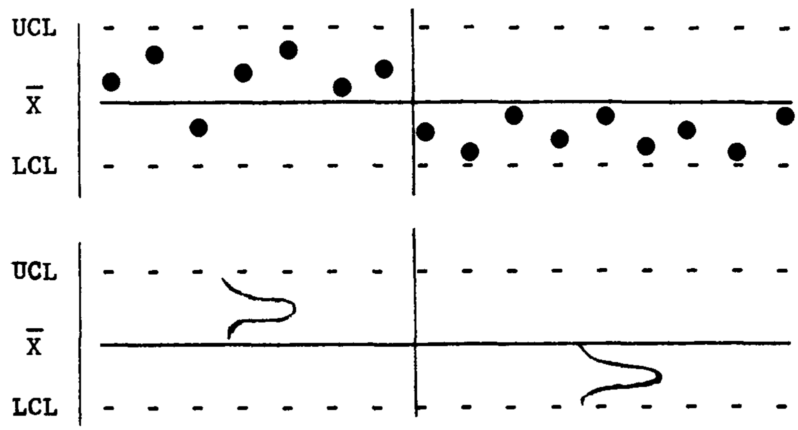
8. No observations near the center value (8 connective values in the 3s zone).



Sampling Periods

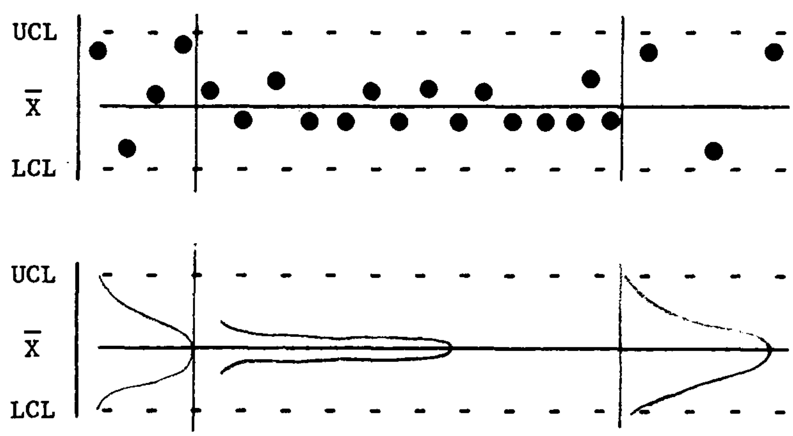
This type of distribution would indicate that the samples are being drawn from two populations instead of from one normal population.

9. Sudden shift in level.



A variable has entered the process at the vertical line and has caused the distribution center to shift (lower in this case).

10. Variability shift



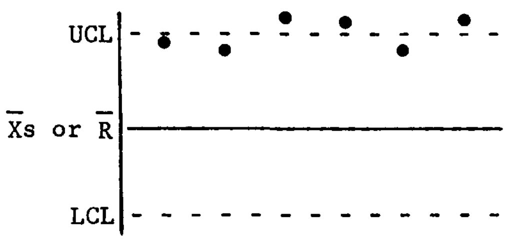
Sampling simultaneous from two different populations.

Using these 3 types of charts:

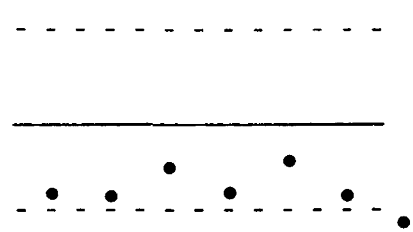
1. Individual value or mean value chart,
2. Standard deviation or range chart,
3. Percent defective,

a production statistic of any nature can be followed on a product and the production can be adjusted to keep the product within control.

11. The standard deviation or range portion of the control chart is usually plotted directly under the individual value chart and the information contained on this chart is equally as important in controlling the operation. Examples are as follows:



Variability is increasing if it cannot be reduced the production mean must be adjusted or a greater percentage of samples will exceed the control limits on the individual value control chart.



Variability is reduced. Production mean may be adjusted and the desired percentage will still be within the control limits on the individual value chart.

Using these 2 types of charts (individual value chart and standard deviation chart) a production statistic of any nature can be followed on a product and the production can be adjusted to keep the product within control.

In some cases the range rather than the standard deviation of the sampling periods is used to monitor the changing variability.

USE OF SPECIFICATIONS

Often maximum or minimum specifications are arbitrarily set by company policy (or regulations), and it is desirable to calculate the percent of the samples that will fall above the maximum specifications or the percent that will fall below the minimum specification. This is accomplished with the aid of a normal curve graph (or table) similar to the one shown in the chapter on "Measures of Variability" and entitled "Area Under a Normal Curve." Steps in calculating are as follows: [the data in the "frankfurter-fat" problem; "$\overline{X}$-R quality control chart" will be used as an example] ; The question to be answered is what percent of the individual observations will exceed 30% fat.

A. Percent of samples between the mean and the specification may be calculated as follows:

$$\text{Number of standard deviations between specification and mean} = \frac{\text{Distance that specification is from the mean}}{s}$$

or

$$= \frac{\text{Distance that specification is from the mean}}{\bar{R}/d_2}$$

Specification = 30%

$\bar{X}$ = 29.7%

$\bar{R}$ = 1.1%

d_2 = 1.7% from graph "Estimating the standard deviation from the Range" and using n = 3

$$\text{Number of standard deviations between specification and mean} = \frac{30 - 29.7}{1.1/1.7} = \frac{0.3}{0.65} = .46$$

Entering the graph entitled "Area Under a Normal Curve" at .46 there are 18% of the observations between the mean (29.7%) and 30%.

B. By substracting the above value from 50% (or adding it to 50 if the specification was below the mean) the percent exceeding 30% can be determined and in this case would be:

% above 30% = 50 - 18 = 32% of observation exceed 30% fat.

C. To calculate the percent below minimum specification, the same procedure is used. In the "frankfurter-fat" example, what percent is below 28% fat?

$$\text{Number of standard deviations between specification and mean} = \frac{29.7 - 28}{1.1/1.7} = \frac{1.7}{0.65} = 2.6$$

Looking under 2.6 in the "Area Under a Normal Curve Graph" gives a value of 49.5% between the mean (29.7) and 28% fat content. The percentage of the samples below this minimum specification (28%) would be 50 - 49.5 or 0.5% of observations below 28% fat.

Cumulative sum (CUSUM) chart.

This type of chart is used when we want to call attention to shifts in levels (Government inspection monitoring technique). It is based on sums of observations rather than on individual observations and a change in quality can be seen more easily by visual inspection.

Definitions

<u>Reference value</u> - this is a constant for each CUSUM chart and this value is subtracted from each sample result (except the value cannot go below zero). It is a type of tolerance.

<u>Action line</u> - line on the CUSUM chart that calls for action when the sample value falls on or exceeds this line (similar to control limits).

<u>Restart line</u> - usually where plotting is started. It is often zero or it can be the same as the rejection restart line.

<u>Rejection restart line</u> - the line on which plotting is started after exceeding the action line.

A pn chart that will also be plotted by the CUSUM method.

Sampling Period	Sample Size n	Number of Defects in Sample h	Fraction Defective $p = \frac{h}{n}$	Number of Defectives $pn = \left(\frac{h}{n}\right) n = h$
A	40	1	.025	1
B	40	1	.025	1
C	40	0	0	0
D	40	1	.025	1
E	40	0	0	0
F	40	3	.075	3
G	40	3	.075	3
H	40	4	.1	4
I	40	3	.075	3
J	40	4	.1	4
	$\Sigma n = 400$	$\Sigma h = 20$		$\Sigma pn = 20$

$$\bar{p} = \frac{\Sigma pn}{\Sigma n} = \frac{20}{400} = .05 \qquad \overline{pn} = \frac{\Sigma pn}{\text{\# of sampling periods}} = \frac{20}{10} = 2$$

99% level

$$UCL_{pn} = \overline{pn} + 3 \sqrt{\overline{pn}\ (1-\bar{p})}$$

$$= 2 + 3 \sqrt{2\ (1-.05)}$$

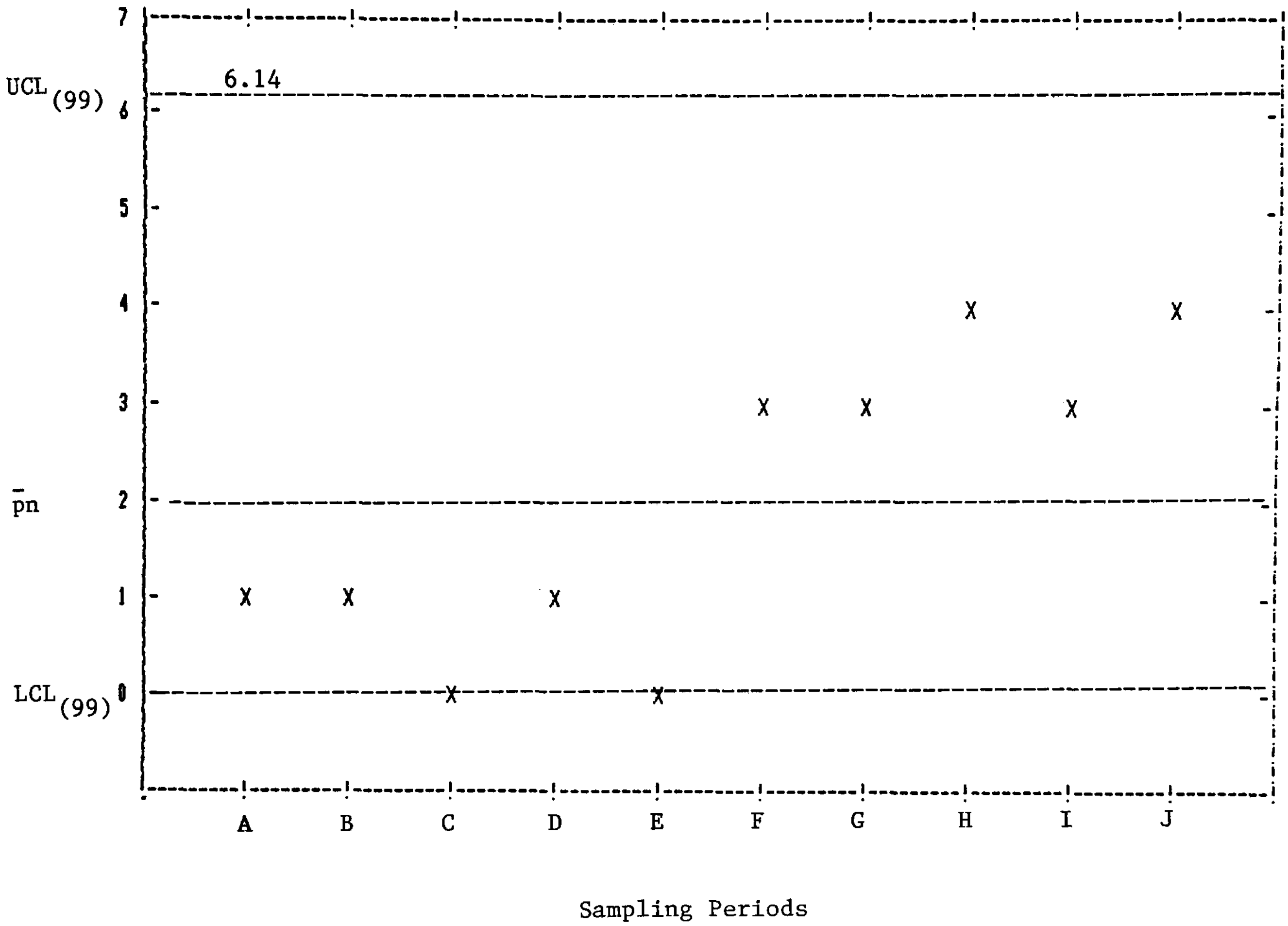
$$= 2 + 4.14$$

$$= 6.14$$

Target value = $\overline{pn}$ = 2

LCL_{pn} = 2 - 4.14 = -2.14 or 0

pn Chart (original data plotted)



Calculation for CUSUM plot:

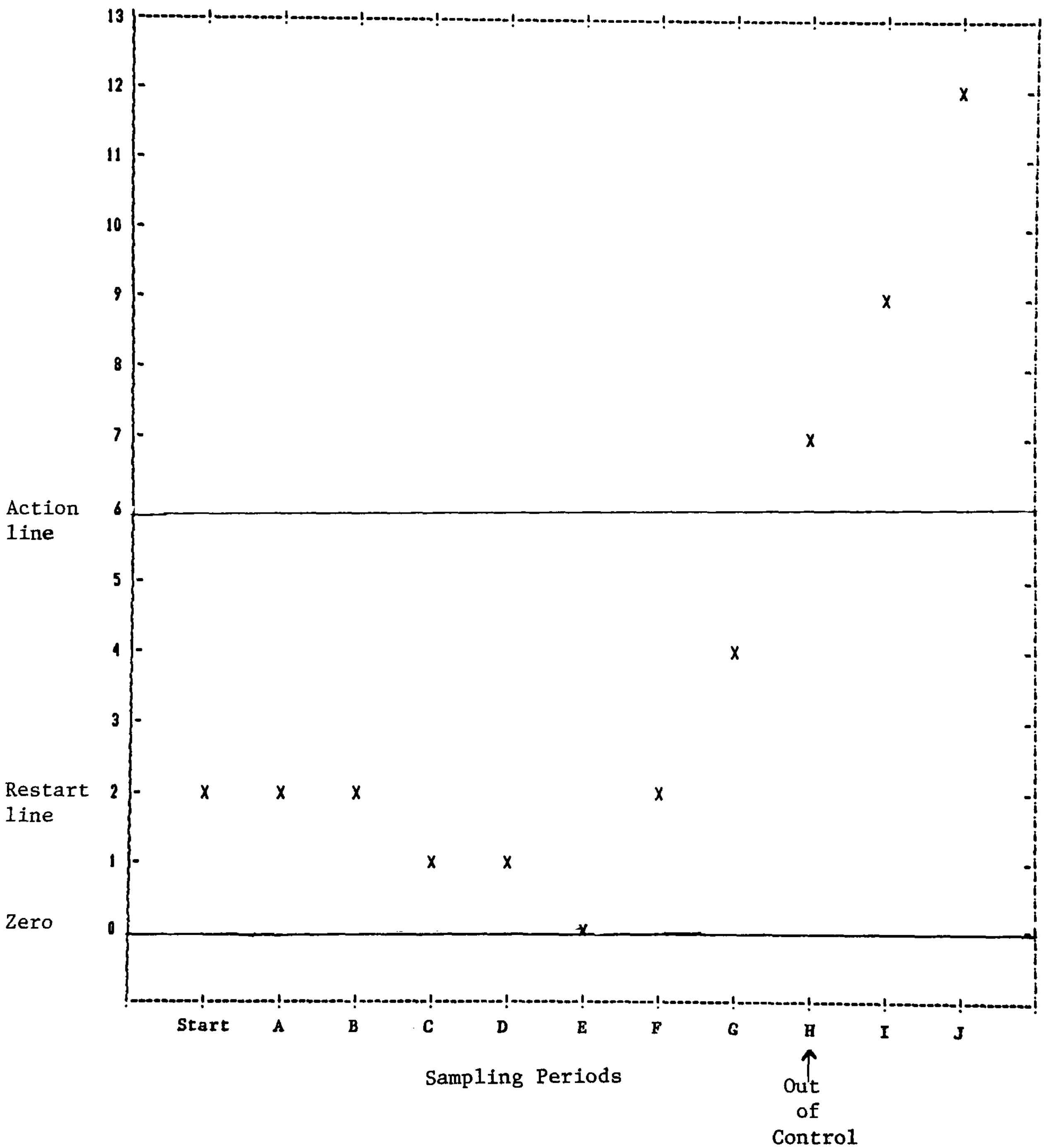
Column A	B	C	D
	Original Observation h	Column B + Previous D	Column C - 1
Start		2	2
A	1	3	2
B	1	3	2
C	0	2	1
D	1	2	1
E	0	1	0
F	3	3	2
G	3	5	4
H	4	8	7
I	3	10	9
J	4	13	12

CUSUM chart of same data

Reference value = 1

Action line = 6

Restart line = 2



References

American Society For Quality Control, State University of Iowa Section. Quality Control Training Manual.

Cowden, D. J. Statistical Methods in Quality Control. Prentice Hall, Inc. Englewood Cliffs, N. J.

Dunstan, W. A. 1972. Personal Communication. Ralston Purina Company. St. Louis, Missouri 63188.

National Canners Association. 1968. Laboratory Manual for Food Canners and Processors, Vol. 2. AVI Publishing, Inc. Westport, Conn.

Naval Ordinance Systems. The Quality Control Chart Technique. U. S. Printing Office, Washington, D. C.

Western Electric Co. Statistical Quality Control. Western Electric Co., Inc. 22 Broadway, N. Y., N. Y.

CORRELATIONS

The previous examples have dealt with one population and in a few cases the comparison of 2 different populations. Correlation and regression (next chapter) analysis are used when there are two observations made on the same (or related) sample. For example, if 100 people were examined for both height and weight, a correlation analysis could be used to determine if there was any relationship between these two (height and weight) variables.

Conditions needed to calculate a Correlation (r is the symbol used to represent a correlation)

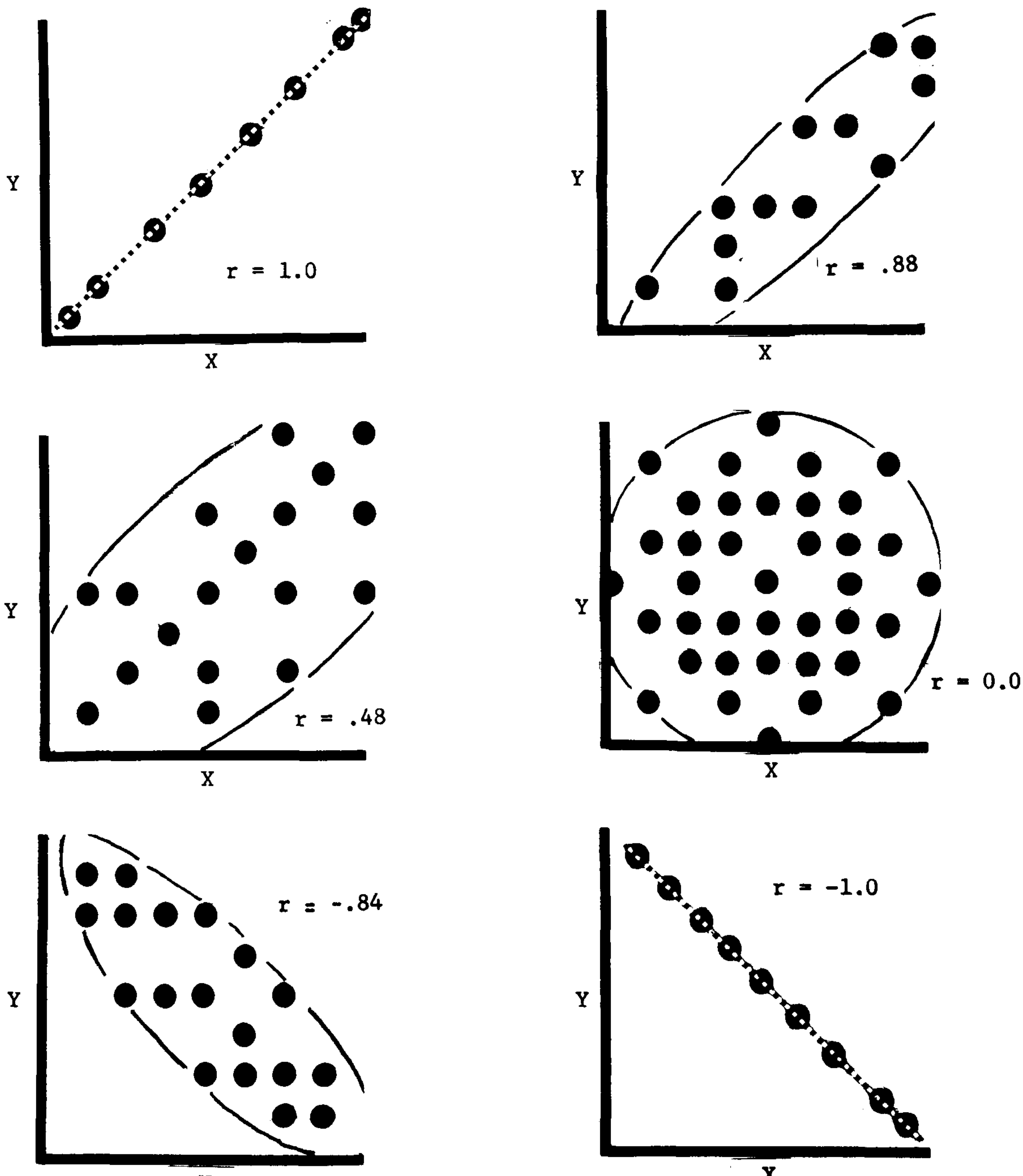
1. Two observations are obtained on each sample (or matched sample) and one is usually designated X and the other Y.
2. It is often desirable to plot on a scatter diagram--Y on the vertical scale and X on the horizontal and visually observe the relationship (or lack thereof) between the two variables. Each point on the diagram represents an X and a Y value.

Properties of a Correlation

1. Correlations range from -1.0, through zero, to +1.0.
2. A correlation of -1.0 indicates a perfect inverse correlation. As X increases Y decreases by an exact amount or as X decreases Y increases by an exact amount.
3. All negative correlations indicate this inverse relationship. The relationship becomes poorer as the correlation approaches zero.
4. Zero correlation indicates no relationship between X and Y.
5. Positive correlations indicate a direct relationship. As X increases Y increases or as X decreases Y decreases. The relationship improves as the correlation approaches plus one.

6. Correlation of +1.0 indicates a perfect direct relationship. As X increases one unit Y increases a given quantity or if X decreases one unit Y decreases a given quantity.

The following scatter diagrams show 6 different correlations ranging from +1.0, +0.88, +0.48, 0.0, -0.84 and -1.0.



If a straight line were fitted to the data notice the following:

1. The closer the data approach the line or the less deviation from a straight line the higher (both positive and negative) the correlation becomes.
2. On positive correlation the line rises as X becomes larger (right end of scale).
3. On negative correlation the line falls as X becomes larger (right end of scale).
4. No line can be fitted to zero correlation. A line in any direction would work about as well as a line in any other direction.

Correlations are calculated as follows:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \quad \sum y^2}} = \frac{\sum (X-\bar{X}) \quad (Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \quad \sum (Y-\bar{Y})^2}}$$

An example for a correlation problem will be the comparison of the thickness of cold cut slices with their weights.

The following are the data to be used in this problem.

1	2	3	4	5	7	8	9	10	12
Sample	Thickness in millimeters X	Average thickness in millimeters $\bar{X} = \frac{\sum X}{n}$	Deviation from the average thickness $x = (X-\bar{X})$	Deviation from the average thickness squared $x^2 = (X-\bar{X})^2$	Weight in grams Y	Average weight in grams $\bar{Y} = \frac{\sum Y}{n}$	Deviation from the average weight $y = (Y-\bar{Y})$	Deviation from the average weight squared $y2 = (Y-\bar{Y})2$	Cross Products xy or $(X-\bar{X})\cdot(Y-\bar{Y})$
A	4	4.0	0	+ 0	33	37.0	- 4.0	+ 16.0	- 0.0
B	2	4.0	-2.0	+4.0	16	37.0	-21.0	+441.0	+42.0
C	3	4.0	-1.0	+1.0	30	37.0	- 7.0	+ 49.0	+ 7.0
D	6	4.0	+2.0	+4.0	62	37.0	+25.0	+625.0	+50.0
E	5	4.0	+1.0	+1.0	44	37.0	+ 7.0	+ 49.0	+ 7.0
n=5	$\sum X=20$		$\sum x=0$	$\sum x^2=10.0$	$\sum Y=185$		$\sum y=0$	$\sum y^2=1180.0$	$\sum xy=106.0$

The first step in evaluating correlation type data is to plot them on a scatter diagram and note any relationship that may exist. The weight in grams will be plotted on the vertical axis and the thickness in millimeters on the horizontal axis.

By observing the scatter diagram the following points may be observed:

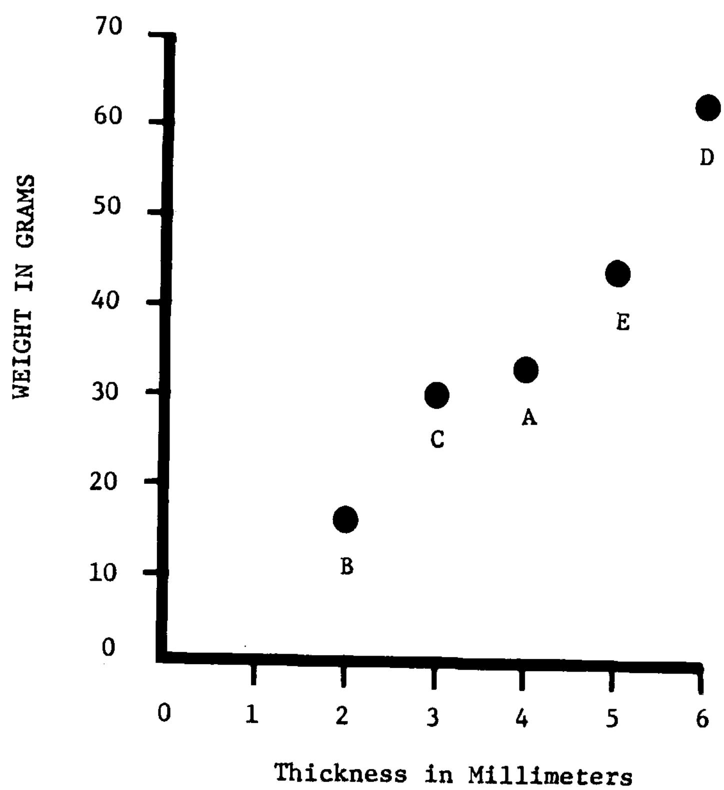
1. The correlation is positive as would be expected.
2. If a straight line was drawn through the data on the diagram there would not be much deviation from this line; therefore, a high positive correlation is expected.

The calculation of this correlation is as follows:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \quad \sum y^2}} = \frac{106}{\sqrt{(10) \quad (1180)}} = \frac{106}{\sqrt{11800}} = \frac{106}{108.62780}$$

$r = +0.9758$ rounded to $+ .98$

This correlation indicates a close positive relationship between the 2 variables (thickness and weight).



It is often easier to calculate correlations without figuring deviations and this may be accomplished on our previous example as follows:

1	2	6	7	11	13
Sample	Thickness in millimeters X	Thickness in millimeters sqd. X^2	Weight in grams Y	Weight in grams sqd. Y^2	Cross Products XY
A	4	16	33	1089	132
B	2	4	16	256	32
C	3	9	30	900	90
D	6	36	62	3844	372
E	5	25	44	1936	220
n=5	$\Sigma X=20$	$\Sigma X^2=90$	$\Sigma Y=185$	$\Sigma Y^2=8025$	$\Sigma XY=846$

$$r = \frac{n\Sigma XY - (\Sigma X)\ (\Sigma Y)}{\sqrt{[n(\Sigma X^2) - (\Sigma X)^2]\ [n(\Sigma Y^2) - (\Sigma Y)^2]}}$$

$$r = \frac{5\ (846) - (20)\ \ (185)}{\sqrt{[5\ (90) - (20)^2]\ [5\ (8025) - (185)^2]}} = \frac{530}{\sqrt{[50]\ [5900]}}$$

$$r = \frac{530}{\sqrt{295000}} = \frac{530}{543.1390} = +0.9758 \quad \text{rounded to } 0.98$$

As can be seen the two calculations yield the same correlation value (r).

If a number of pounds of cold cuts had been examined and the average thickness of a slice (X) in the pound had been compared to the number of slices required to weigh a pound (Y), a negative correlation would have been obtained.

If the correlation value is squared (r^2) and then multiplied by 100 this product will indicate the percentage of the total variation in Y that is accounted for by X.

$$\text{\% of variation in Y accounted for by X} = r^2\ (100)$$

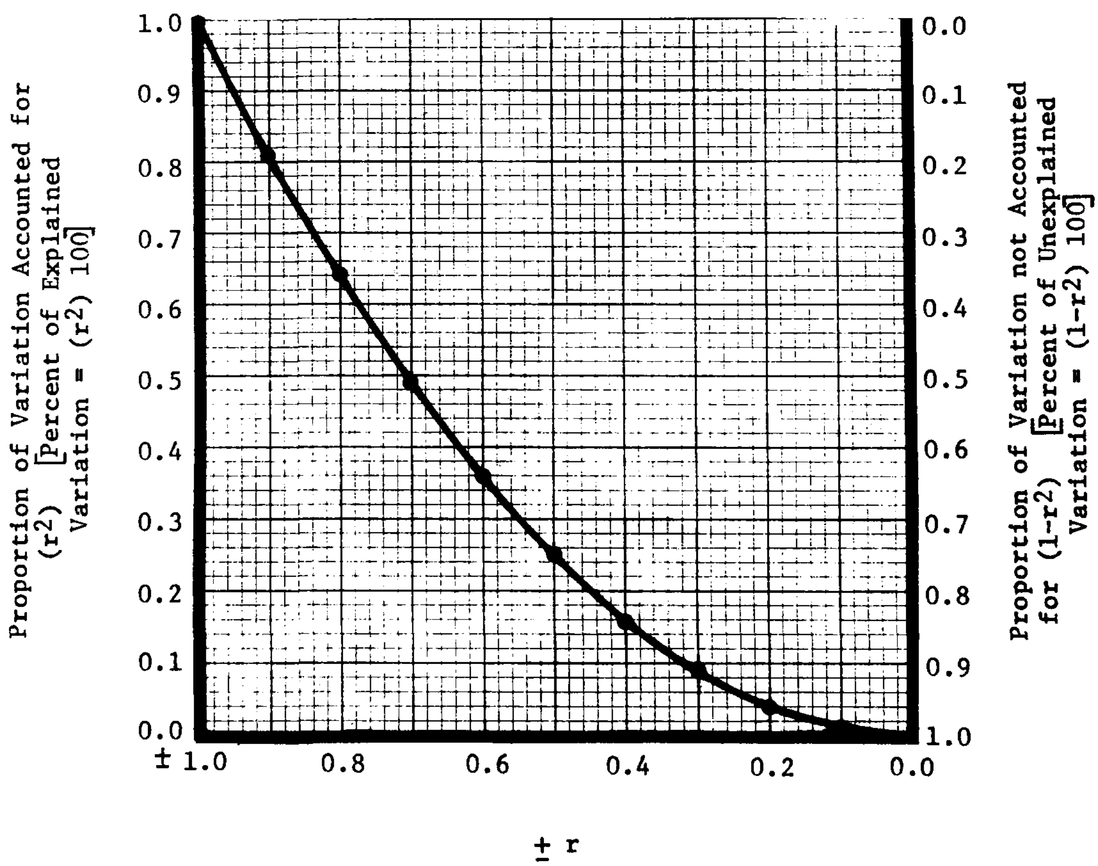
In the previous example:

$$\begin{aligned}\text{\% of variation in Y accounted for by X} &= (.98)^2(100)\\ &= .96\ (100)\\ &= 96\%\end{aligned}$$

This indicates that 96% of the variation in sample weight (in grams) is accounted for by the thickness of the sample (in millimeters). This indicates that 4% (100-96%) of the variation in weight is not accounted for by thickness and must be attributed to other factors (an example would be density of the product).

The relationship between correlation values and the preparation of the variability accounted for (or not accounted for) can be seen in the following graph.

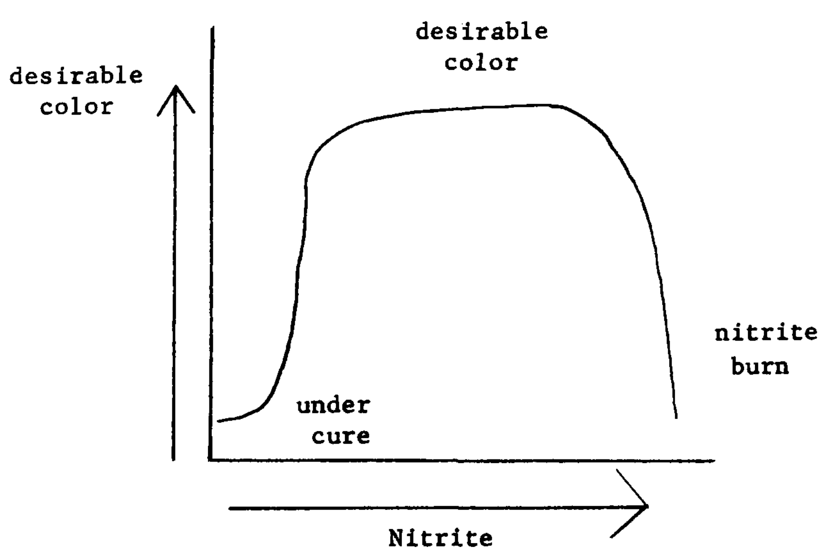
Proportion of Variation Accounted for (or not Accounted for) at Various Correlation (r) Values.



Errors of Interpretation

Often correlation values are misinterpreted and the following danger signs will aid in avoiding these pitfalls.

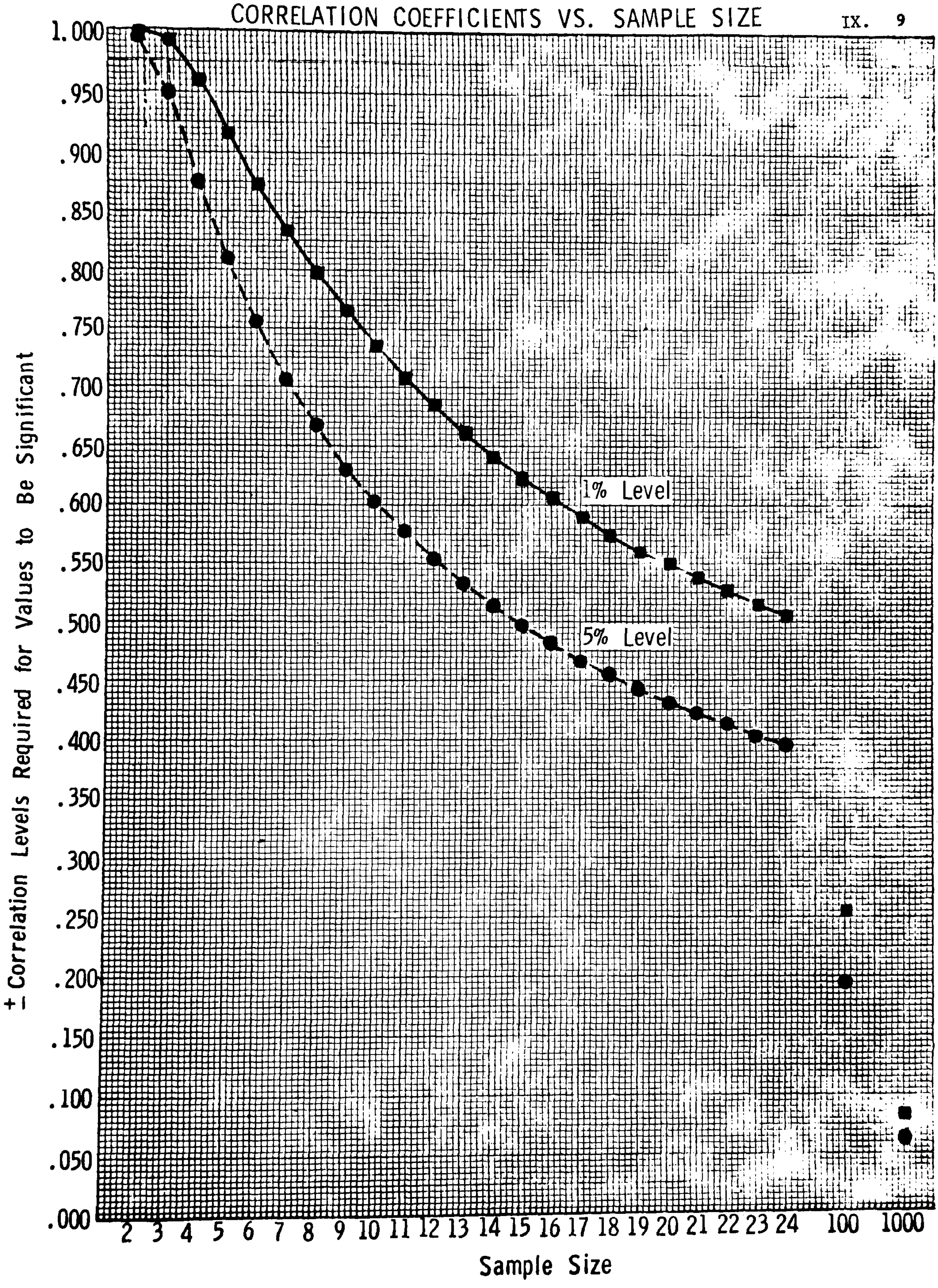
1. Correlation values yield no information as to cause and effect relationship. In the example used increased thickness was responsible for increased weight but reasoning rather than correlation values supplied this information. If a correlation analysis were run on the weight of slaughter steers (X) and the weights of radios sold (Y) using data from the last 20 years there would be a positive correlation, (both weights decreased) but certainly one did not cause the other. They are both related to time.
2. Correlation values are high only if the relationship is linear (straight line) and may be quite low when the relation is curvilinear (curved line) even though the relationship may be quite good. This type of relationship can often be spotted when the scatter diagram is plotted. This type of relationship would be obtained if desirable color scores (Y) were plotted against nitrite concentration (X) of a sausage product.



In this example the correlation value would be low (relationship non-linear) but this does not mean there is no relationship between nitrite level and cured color.

3. Restricting the range of one of the variables will often yield a reduced correlation value. For example: If more values in the previous example had been between 3 and 4 millimeters and the ranges had been limited between these 2 values, a lower correlation coefficient would have been obtained and the false impression may have been reached if the conclusion had been that there was little relationship between thickness and weight.

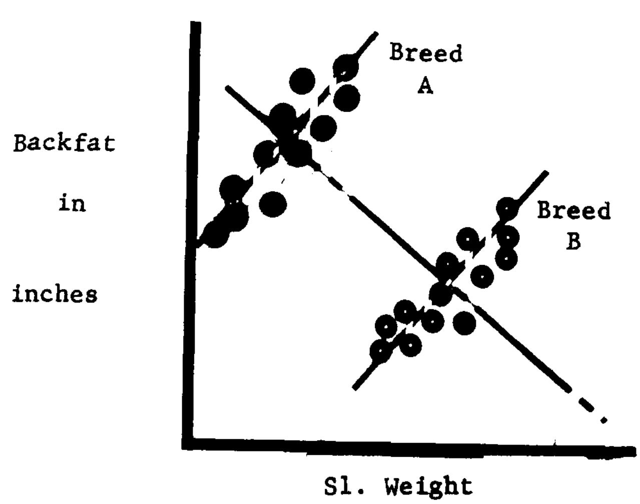
4. Effect of Number: The number of observations does not influence the size of the calculated correlation value but does influence how accurately the relationship between the two variables is measured. For this reason, the smaller the sample size the larger the correlation coefficient must be to achieve a significant value. Tables are available that indicate the exact size of the correlation coefficient required for significance with different sample sizes. The relationship between sample size and magnitude of correlation value required for significance can be seen in the following graph.


CORRELATION COEFFICIENTS VS. SAMPLE SIZE
± Correlation Levels Required for Values to Be Significant
1.000
.950
.900
.850
.800
.750
.700
.650
.600
.550
.500
.450
.400
.350
.300
.250
.200
.150
.100
.050
.000
1% Level
5% Level
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 100 1000
Sample Size

In the previous "cold cut" problem it requires a correlation value of 0.81 (5% level of significance) to be significant and 0.92 (1% level of significance) to be highly significant. This means that if there were no relationship between the X and Y values and if 5 pairs of X and Y values were sampled 100 different times from this population that a correlation value of 0.81 or greater would be expected five times by chance alone. Under these same conditions a correlation value of 0.92 would be expected one time out of a hundred by chance alone.

Since a correlation value of 0.98 was obtained from the "cold cut" experiment a highly significant (1% level) relationship exists between sample thickness and sample weight unless, by chance alone, the one sample out of 100 that is expected to have a correlation value in excess of 0.92 has been selected (highly unlikely).

5. Samples containing subgroups with unequal means: Correlation should not be calculated on samples which contain subgroups with unequal means, because the groups chosen have more influence on the correlation than the relationship of the two variables. The following example will illustrate this fallacy.



In both breed A and B the relationship between slaughter weight and backfat thickness is positive but if the 2 groups are combined in one sample, a negative correlation would result.

Sample Problems

1. The following information was obtained from 4 members of the class.

Class Member	Shoe Size	Hat Size
A	9	7
B	10	8
C	9	6
D	8	7

Plot the above information on a scatter diagram and estimate the correlation value. Calculate the correlation value and check your results. Is this a significant correlation?

(Answer r = .5, No)

2. The following information was obtained from 4 members of the class.

Class Member	Shirt Neck Size	Waist Size
A	15	32
B	18	40
C	16	34
D	15	34

Plot the above information on a scatter diagram and estimate the correlation value. Calculate the correlation value and check your results. Is this a significant correlation? At what level?

(Answer r = +.95, Yes, between 95 and 99% level)

References

Bartz, Albert. 1965. Elementary Statistical Methods for Educational Measurement. Burgess Publishing Company. Minneapolis, Minnesota.

Chou, Ya-lun. 1965. Applied Business and Economic Statistics. Holt, Rinehart and Winston. New York, Chicago, San Francisco, Toronto, London.

Haber, Audrey and Richard P. Runyon. 1969. General Statistics. Addison-Wesley Publishing Company, Reading, Massachusetts; Menlo Park, California; London; Don Mills, Ontario.

Snedecor, George W. 1959. Statistical Methods. The Iowa State College Press, Ames, Iowa.

Walker, Helen M. and Joseph Lev. 1958. Elementary Statistical Methods. Holt, Rinehart and Winston, New York.

REGRESSION

Two observations are made on each of several samples (or related samples) and these data are then plotted on a scatter diagram. Correlation analysis performed on the data indicates if there is any relationship between the variables. The next task is to draw a line on the scatter diagram that best shows the relationship between the two variables. This line is often fitted by eye and drawn freehand on the diagram. A respectable job may often be accomplished by this crude technique but the "best fitting line" must be calculated, a formula determined, and the line drawn according to the formula. After the formula is determined it is possible to calculate the expected value of one variable if the other variable is known. The calculation of the "best fitting line" is called regression analysis. Both straight and curved lines may be fitted to data and the same principles are involved but the more complex the curve the more difficult the calculation becomes and only straight line fitting will be covered in this text.

With correlation analysis it made no difference which variable was called X and which was called Y but in regression analysis it does.

It is customary to use the following descriptions:

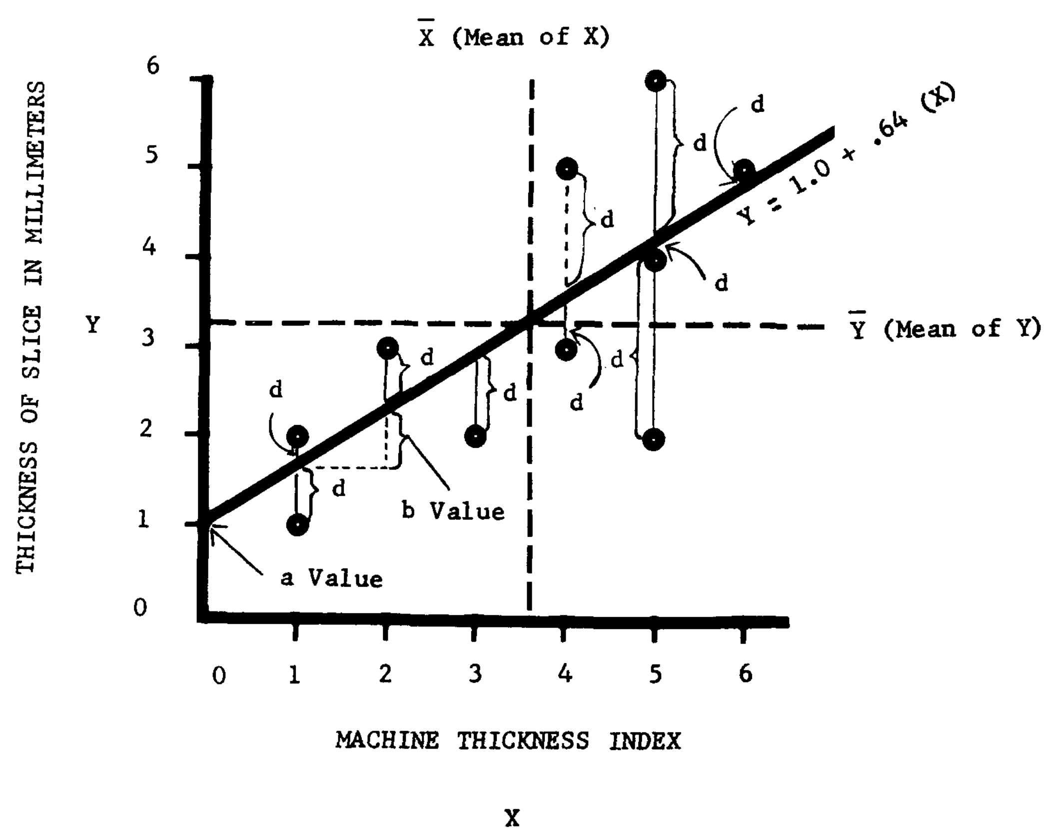
X - independent variable -- values are known -- plotted on horizontal axis

Y - dependent variable -- values are to be determined -- plotted on the vertical axis

With the above designations a relationship is determined between X and Y, a formula is developed that best fits the data, and then Y can be predicted from X values. X values cannot be predicted from Y values using this formula.

What is meant by the line that best fits the data? The following scatter graph will illustrate the principles used. This graph shows data obtained from an experiment using a very poor slicing machine; X is the thickness index setting on the machine (horizontal axis-independent variable - may be dialed to desired setting).

The value Y is the actual thickness of the meat slice in millimeters. (Vertical axis-dependent variable-depends primarily on the X setting.)



The "best line" is calculated so that the sum of the deviation from the line (labeled d on graph) squared will be smaller than if any other straight line had been used. This same principle was used in the calculation of the mean. The line also crosses the data at the same point that the average X (average machine setting) and average Y (average thickness of a slice) values cross. Any other straight line that could be drawn would result in the $\sum d^2$ values being larger. The equation for a straight line is:

$Y = a + b X$	a & b	are constants for the data and may be calculated from them
	X & Y	are variables

For the example used the a and b values (an explanation of how they were calculated in this problem is given later) are

$$a = +1.0$$

$$b = +0.64$$

and thus the equation for these data becomes

$$Y = a + bX$$

$$Y = 1.0 + 0.64\ (X)$$

The "a" value, or 1.0 in this case, is the point where the line will cross the Y axis when X is zero. The meat slice will be 1 millimeter thick if the index is set on zero.

The "b" value or +.64 in this case is the slope of the line. Each time the line increases 1 unit in X it increases .64 units in Y. For example, the value for Y, at X equals 1, is 1.64. The value for Y, at X equals 2, is 2.28. The difference in Y for an increase of one in X (from 1 to 2) is 2.28 minus 1.64 or .64 which is the "b" value or the slope of the line. This means that the slice will get .64 of a millimeter thicker each time we move the index up one unit.

The regression line thus becomes

$$Y = 1.0 + 0.64X$$

The comparison of the predicted Y Value ($\hat{Y}$) to the actual Y Value can be seen as follows:

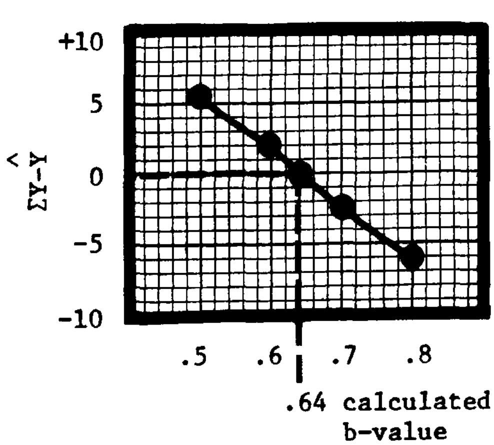
Sample	Machine setting X	Thickness in millimeters Y	$Y=.9929577464+0.6408450704X$ $\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
A	1	1	1.633802816	-.633802816	.401706010
B	1	2	1.633802816	.366197184	.134100378
C	2	3	2.274647887	.725352113	.526135688
D	3	2	2.915492957	-.915492957	.838127354
E	4	3	3.556338028	-.556338028	.309512001
F	4	5	3.556338028	1.443661972	2.084159889
G	5	2	4.197183098	-2.197183098	4.827613566
H	5	4	4.197183098	-.197183098	.038881174
I	5	6	4.197183098	1.802816902	3.250148782
J	6	5	4.838028168	.161971832	.026234874
				Σ+.000000006 or 0	Σ12.436619716 or 12.44

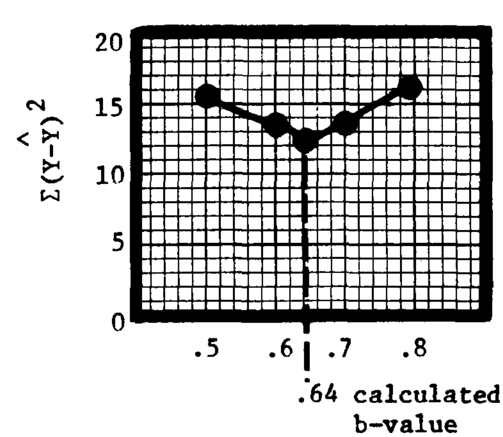
The sum of the deviations around the regression equal zero and the sum of these deviations squared equal the smallest possible number.

If the slope had been different from 0.6408450704 then the deviation should not equal 0 and the deviation squared should be more than 12.4. To prove this point slopes 0.5, 0.6, 0.7, and 0.8 will be compared.

			b=0.5			b=0.6			b=0.7			b=0.8		
Sample	X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
A	1	1	1.49	-.49	.24	1.59	-.59	.35	1.69	-.69	.47	1.79	-.79	.62
B	1	2	1.49	.51	.26	1.59	.41	.17	1.69	.31	.10	1.79	.21	.04
C	2	3	1.99	1.01	1.02	2.19	.81	.66	2.39	.61	.37	2.59	.41	.17
D	3	2	2.49	-.49	.24	2.79	-.79	.62	3.09	-1.09	1.19	3.39	-1.39	1.93
E	4	3	2.99	.01	.00	3.39	-.39	.15	3.79	-.79	.62	4.19	-1.19	1.42
F	4	5	2.99	2.01	4.04	3.39	1.61	2.59	3.79	1.21	1.46	4.19	.81	.66
G	5	2	3.49	-1.49	2.22	3.99	-1.99	3.96	4.49	-2.49	6.20	4.99	-2.99	8.94
H	5	4	3.49	.51	.26	3.99	.01	.00	4.49	-.49	.24	4.99	-.99	.98
I	5	6	3.49	2.51	6.30	3.99	2.01	4.04	4.49	1.51	2.28	4.99	1.01	1.02
J	6	5	3.99	1.01	1.02	4.59	.41	.17	5.90	-.90	.81	5.79	-.79	.62
Σ				+5.10	15.60		+1.09	12.71		-2.81	13.74		-5.70	16.40

The deviation and the deviation squared for these slopes can be graphed as follows.





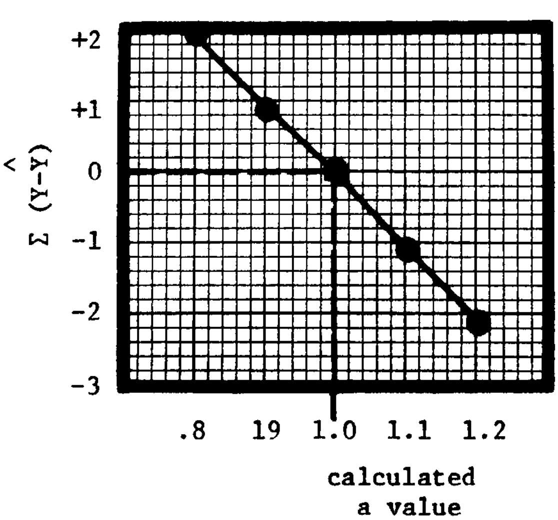
This illustrates as the slope changes the deviations change and passes through zero when the appropriate b value is graphed.

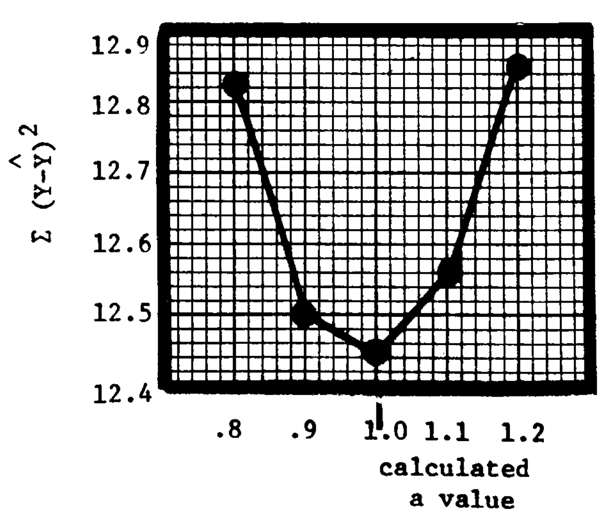
The "deviation squared" decreases as the appropriate b value is approached and reaches a minimum value when it is reached. The "deviation squared" value increases as the b values are changed beyond the appropriate b value.

If the intersection (a value) had been different from 0.9929577464 then the deviation would not equal zero and the deviation squared would be more than 12.4. To prove this point intersections of 0.8, 0.9, 1.1 and 1.2 will be compared.

			a=0.8			a=0.9			a=1.1			a=1.2		
Sample	X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
A	1	1	1.44	-.44	.19	1.54	-.54	.29	1.74	-.74	.55	1.84	-.84	.71
B	1	2	1.44	.56	.31	1.54	.46	.21	1.74	.26	.07	1.84	.16	.03
C	2	3	2.08	.92	.85	2.18	.82	.67	2.38	.62	.38	2.48	.52	.27
D	3	2	2.72	-.72	.52	2.82	-.82	.67	3.02	-1.02	1.04	3.12	-1.12	1.25
E	4	3	3.36	-.36	.13	3.46	-.46	.21	3.66	-.66	.44	3.76	-.76	.58
F	4	5	3.36	1.64	2.69	3.46	1.54	2.37	3.66	1.34	1.80	3.76	1.24	1.54
G	5	2	4.00	-2.00	4.00	4.10	-2.10	4.41	4.30	-2.30	5.29	4.40	-2.40	5.76
H	5	4	4.00	.00	.00	4.10	-.10	.01	4.30	-.30	.09	4.40	-.40	.16
I	5	6	4.00	2.00	4.00	4.10	1.90	3.61	4.30	1.70	2.89	4.40	1.60	2.56
J	6	5	4.65	.36	.13	4.75	.25	.06	4.95	.05	.00	5.05	-.05	.00
Σ				+1.96	12.82		+.95	12.51		-1.05	12.55		-2.05	12.86

The deviation and the deviation squared for these intercepts can be graphed as follows.





This illustrates as the intercept ("a" value) changes the deviations change and pass through zero when the appropriate "a" value is graphed.

The "deviation squared" decrease as the appropriate "a" value is approached and reaches a minimum value when it is reached. The "deviation squared" value increases as the "a" values are changed beyond the appropriate "a" value.

Both the a and b values are labeled on the graph.

The calculation of the regression line from the data plotted on the graph is as follows:

Sample	Machine Setting X	X^2	Thickness in millimeters Y	Y^2	XY
A	1	1	1	1	1
B	1	1	2	4	2
C	2	4	3	9	6
D	3	9	2	4	6
E	4	16	3	9	12
F	4	16	5	25	20
G	5	25	2	4	10
H	5	25	4	16	20
I	5	25	6	36	30
J	6	36	5	25	30
	$\sum X=36$	$\sum X^2=158$	$\sum Y=33$	$\sum Y^2=133$	$\sum XY=137$

$n = 10$

$$\bar{X} = \frac{\sum X}{n} = \frac{36}{10} = 3.6$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{33}{10} = 3.3$$

The formula for determining the slope or the "b" value is as follows:

$$b = \frac{n \sum XY - (\sum X) \quad (\sum Y)}{n \sum X^2 - (\sum X)^2}$$

Using the above example

$$b = \frac{10(137) - (36) \quad (33)}{10(158) - (36)^2} = \frac{182}{284} = +0.6408450704 \text{ or } +0.64$$

The formula for determining the intercept or "a" value is

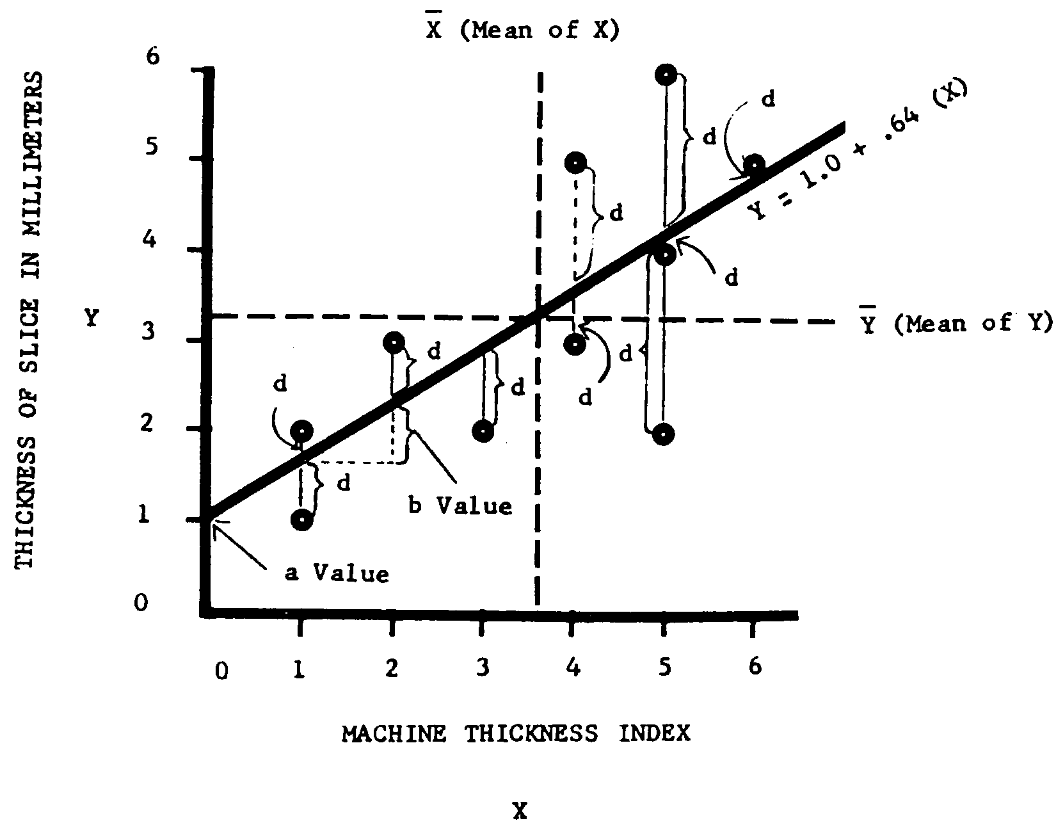
$$a = \bar{Y} - b\,\bar{X}$$

Using the above example

$$a = 3.3 - .6408450704\ (3.6)$$

$$a = 3.3 - 2.307042253$$

$$a = +.9929577464 \text{ or } +1.0$$


X̄ (Mean of X)
Ȳ (Mean of Y)
Y = 1.0 + .64 (X)
THICKNESS OF SLICE IN MILLIMETERS
Y
MACHINE THICKNESS INDEX
X
a Value
b Value
d
0
1
2
3
4
5
6

To plot the regression line on a graph any 2 values are assigned to X and the corresponding Y value is determined.

Example:

X	Y
1	1.64
5	4.20

X = 1 (machine setting of one)

$Y = 1.0 + .64(1)$

$Y = 1.64$ (slice thickness with machine setting of one)

X = 5 (machine setting of five)

$Y = 1.0 + .64(5)$

$Y = 4.20$ (slice thickness with machine setting of five)

The 2 points (X=1, Y=1.64) (X=5, Y=4.20) are plotted on the graph and connected with a line, which is a linear regression line.

For any value of X (machine setting) between 0 and 6 (range of the original data) it is now possible to predict a most likely Y value (thickness) by simply substituting X in the regression formula.

If the correlation between X and Y is perfect (-1 or +1) the same regression line would result regardless of which value was designated -- the independent or dependent variable. However, as the correlation becomes poorer (moves from -1 or +1 toward 0) two lines would be obtained if the dependent and independent variables are reversed. When the correlation reaches 0 the two lines would cross at right angles.

Normally only one equation (the one previously described) is used since usually it is only practical to predict the dependent variable (thickness from the machine setting). However, in a few cases, it makes about as much sense to predict X as Y. In the above example this would be predicting the machine setting from the thickness of the slice and this would rarely be done but in other examples it might be desirable.

In the case where X is to be predicted from Y values the "best equation" is the line that makes the sum of deviation from the line (in the horizontal direction <u>rather than</u> the vertical direction) squared as small as possible. In our previous example the following formula would be used for this type of a prediction.

"b" for predicting $X = \dfrac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2} = \dfrac{182}{10(133) - (33)^2} = \dfrac{182}{241}$

$= +.76$

"a" for predicting $X = \overline{X} - b\,\overline{Y}$

$= 3.6 - .76(3.3) = 3.6 - 2.51$

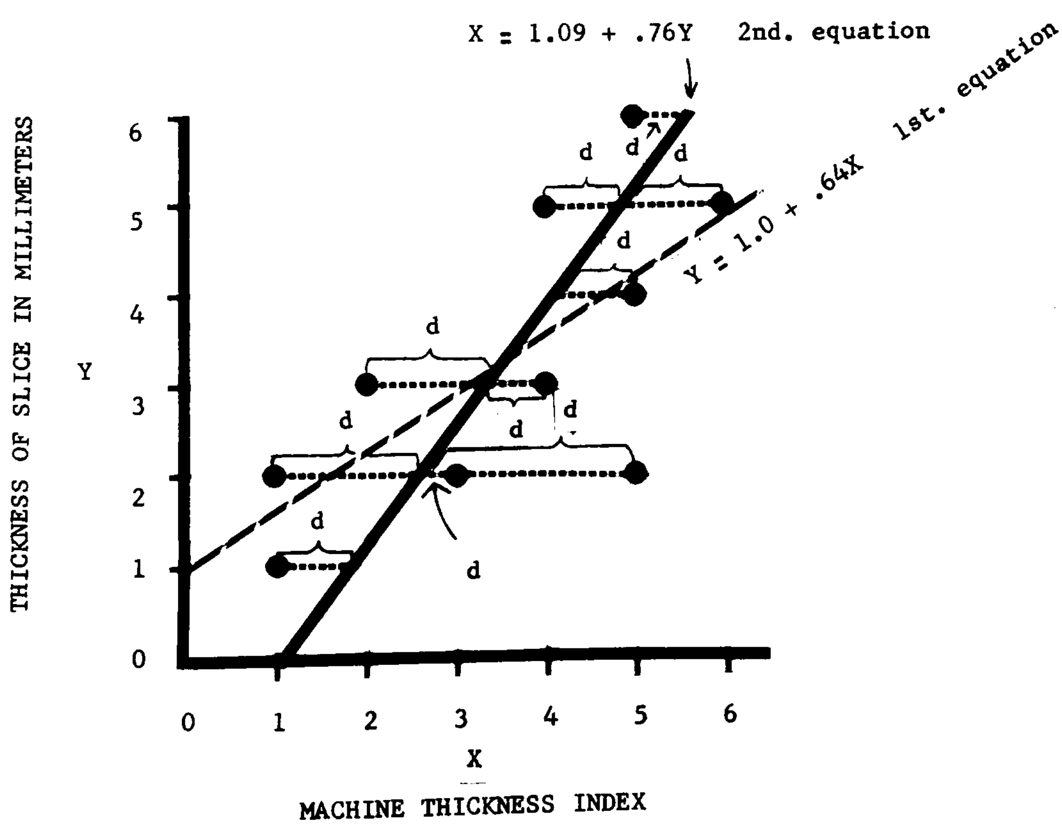
$= 1.09$

This gives the equation:

$$X = a + bY$$

$$X = 1.09 + .76(Y)$$

If both lines were plotted on the original data the graph would appear as follows:



Again the line passes through the point of intersection of the mean of X and the mean of Y. The value "a" in this case is the value of X when Y equals 0. The "b" value is the change in X of a one unit change in Y.

Standard error of estimate ($s_{Y \cdot X}$)

Often it is desirable to know the deviations around the regression line so that an expected maximum or minimum may be used in an estimate of Y rather than the average value. This expected variation is determined by the standard error of estimate and this term has many of the same properties as the standard deviation discussed in the section describing the mean.

For the formula (predicting Y from X)

$$Y = a + bX$$

The standard error of estimate may be calculated from the correlation results as follows:

$$s_{Y \cdot X} = s_Y \sqrt{1 - r^2}$$

s_Y = standard deviation for the Y variable

For the example in this chapter: calculations not shown.

$$r = 0.69$$

$$s_Y = 1.55$$

$$s_{Y \cdot X} = 1.55 \sqrt{1 - (.69)^2}$$

$$= 1.55 \sqrt{1 - .47}$$

$$= 1.55 \sqrt{.53}$$

$$= 1.55 \times .72$$

$$= 1.11$$

The standard error of estimate can also be calculated from the raw data as follows:

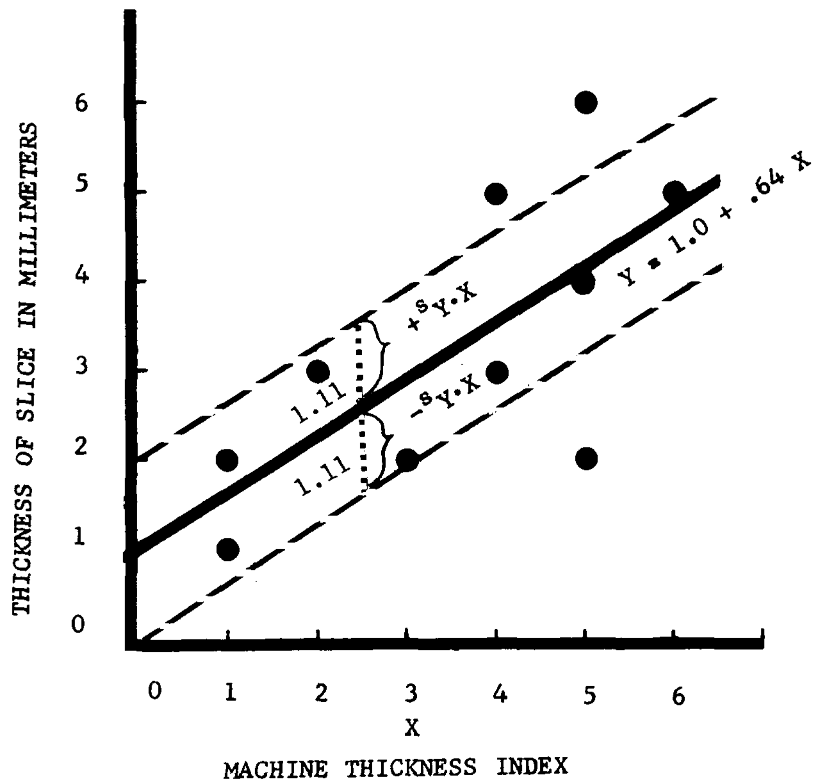
$$s_{Y \cdot X} = \sqrt{\frac{\Sigma Y^2 - a\ \Sigma Y - b\ \Sigma XY}{n}}$$

For the example in this chapter:

$$s_{Y \cdot X} = \sqrt{\frac{133 - 1(33) - .64(137)}{10}} = \sqrt{\frac{12.32}{10}}$$

$$= \sqrt{1.23} = 1.11$$

If a vertical distance of $\pm$ 1 $^{s}Y{\cdot}X$ is drawn on each side of the regression line it should enclose 68% of the data plotted for a large sample. A distance of $\pm$ 2 $^{s}Y{\cdot}X$ and $\pm$ 3 $^{s}Y{\cdot}X$ should enclose 95 and 99.7% respectively. The previous example is repeated with the dashed lines indicating $\pm$ 1 $^{s}Y{\cdot}X$.



In this example $\pm$ 1 $^{s}Y{\cdot}X$ encloses 7 out of 10 observations for 70% of the total but in a larger sample we would expect only 68% to fall in this range.

For any given value of X (in our range) it is possible to predict a value for Y by the following:

$$Y = a + bX$$

The calculated value of Y would be the most likely value (average value). If a range rather than a specific value was desired the range would be between

$$Y \pm 1\ {}^{s}Y{\cdot}X$$

The actual value of Y should be within this range 68% of the time.

ERROR OF INTERPRETATION

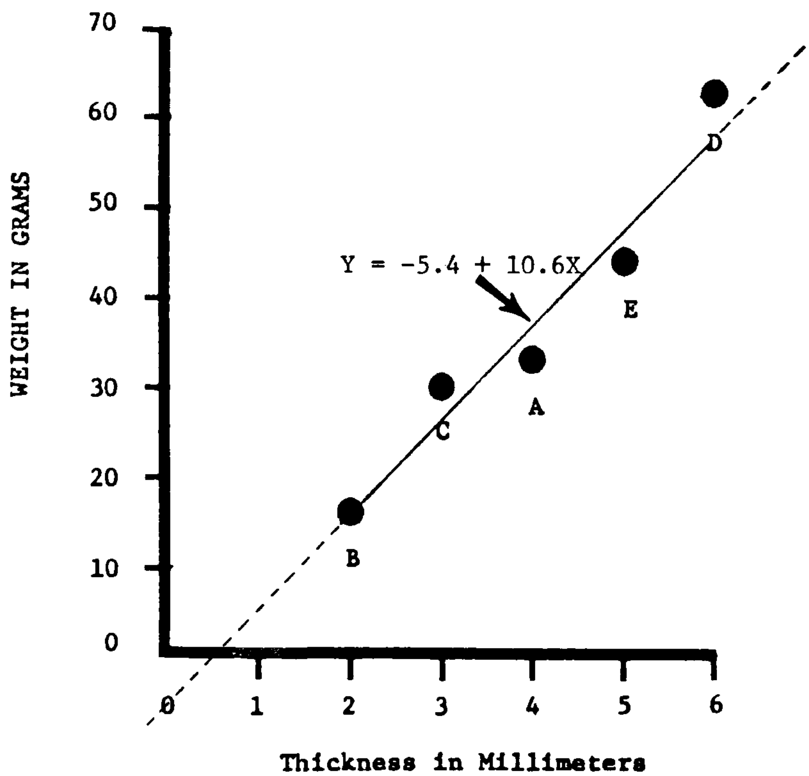
The most tempting error of interpretation in regression analysis is to extend the regression line beyond the areas where the original data were obtained. This is particularly dangerous if the relationship is not completely linear (straight line) over the total range that you are trying to predict (quite frequently the case).

For example, if a regression equation was calculated for the "cold cut -- thickness vs. weight problem" of Chapter IX (not machine setting vs thickness of this chapter) the following results would be obtained:

$$b = \frac{n\Sigma XY - (\Sigma X)\ (\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{530}{30} = 10.6$$

$$a = \overline{Y} - b\overline{X} = 37 - (10.6)(4) = -5.4$$

$$Y = -5.4 + 10.6X$$



Data from this problem range from X = 2 millimeters to X = 6 millimeters in thickness. If information was needed on a thicker or thinner slice it would be tempting to use this regression formula, however, there are no assurances that the formula is accurate outside the range of the original data. For example, the regression formula suggests that a cold cut of zero thickness would weigh -5.4 grams. If a thickness of 8 millimeters was placed in the regression formula the estimated weight might be equally as much in error.

Sample Problems

1. Calculate the regression line on the previously worked correlation problem -- Chapter VIII - Problem 1. (Most of the initial calculation is completed). Predict hat size from shoe size (shoe size is X values).

Class Member	X Shoe size	Y Hat size
A	9	7
B	10	8
C	9	6
D	8	7

A. Plot the above information on a scatter diagram and estimate the a and b values.

B. Calculate the b and a value.

C. Draw the regression line.

D. Calculate the standard error of estimate.

E. Draw lines that are + or - one standard error of estimate from the regression line.

F. What percent of the original data fall between these lines? What percent was expected?

Answers (B - +.5 and +2.5, D - .61, F - 75% and 68%)

2. The following observations were obtained.

X	Y
1	3
3	3
5	6

A. Calculate the mean for X and check the answer.
B. Calculate the mean for Y and check the answer.
C. Calculate the median for X and check the answer.
D. Calculate the median for Y and check the answer.
E. Calculate the mode for X and Y.
F. Calculate the range for X.
G. Calculate the range for Y.
H. Calculate the mean deviation for X.
I. Calculate the mean deviation for Y.
J. Calculate the variance for X and check the answer.
K. Calculate the variance for Y and check the answer.
L. Calculate the standard deviation for X and check the square root.
M. Calculate the standard deviation for Y and check the square root.
N. Calculate the coefficient of variation for X.
O. Calculate the coefficient of variation for Y.
P. Does X or Y have the most variation?
Q. Plot the above data on a scatter diagram and estimate the correlation value.
R. Calculate the correlation. How does it agree with the estimate?
S. Estimate the regression formula.
T. Calculate the "b" value.
U. Calculate the "a" value.
V. Draw the regression line on the scatter diagram.
W. Calculate the "standard error of estimate".
X. Plot 1 standard error of estimate on each side of the regression line. What percentage of the original observation would you expect to fall between these lines?

Answers (A-3, B-4, C-3, D-3.25, E-none and 3, F-4, G-3, H-1.33, I-1.33, J-2.67, K-2, L-1.63, M-1.41, N-54.33%, O=35.25%, P-X, R-(+.86), T-(+.75), U-(+1.75), W-.71, X-68%)

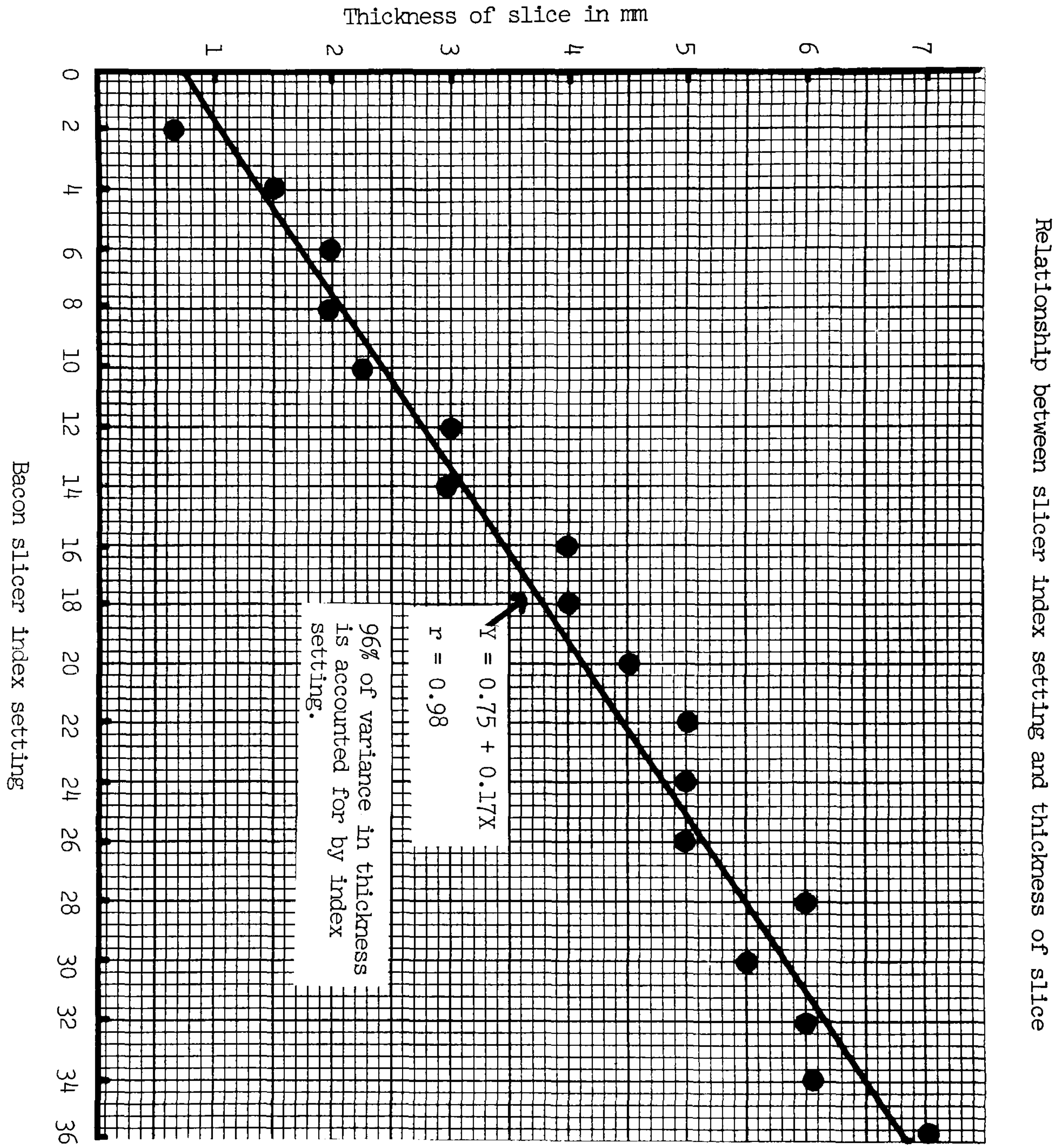
3. The following observations were obtained.

X	Y
1	1
3	2
5	3

A. Calculate the mean for X and check the answer.
B. Calculate the mean for Y and check the answer.
C. Calculate the median for X and check the answer.
D. Calculate the median for Y and check the answer.
E. Calculate the mode for X and Y.
F. Calculate the range for X.
G. Calculate the range for Y.
H. Calculate the mean deviation for X.
I. Calculate the mean deviation for Y.
J. Calculate the variance for X and check the answer.
K. Calculate the variance for Y and check the answer.
L. Calculate the standard deviation for X and check the square root.
M. Calculate the standard deviation for Y and check the square root.
N. Calculate the coefficient of variation for X.
O. Calculate the coefficient of variation for Y.
P. Does X or Y have the most variation?
Q. Plot the above data on a scatter diagram and estimate the correlation value.
R. Calculate the correlation. How does it agree with the estimate?
S. Estimate the regression formula.
T. Calculate the "b" value.
U. Calculate the "a" value.
V. Draw the regression line on the scatter diagram.
W. Calculate the "standard error of estimate".
X. What does the result of "w" indicate?

Answers (A-3, B-2, C-3, D-2, E-no mode, F-4, G-2, H-1.33, I-.67, J-2.67, K-.67, L-1.63, M-.82, N-54.33%, O-41%, P-X, R-+1.0, T-(+.5), U-(+.5), W-0, X-all values are on the line).

Data Taken From Bacon Slicer



References

Spiegel, Murray R. 1961. Theory and Problems of Statistics. McGraw-Hill Book Company, New York, St. Louis, San Francisco, Toronto, Sydney.

Walker, Helen and Joseph Lev. 1958. Elementary Statistical Methods. Holt, Rinehart and Winston, New York.

SIGNIFICANCE

In any study of statistics the word significance will be encountered and the understanding of this term is essential to comprehending statistics.

An experienced hog buyer looks at a pen of slaughter hogs and estimates the average weight as 220 pounds. If the actual average weight of this pen of hogs is 220 pounds then the hog buyer has made an extremely lucky prediction. By knowing this man's skill and by stating a range for example -- 210 to 230 pounds (range will depend on his skill), it is possible to state that there is a 68% chance of the actual average weight of the pigs being in this estimated range. If the range is increased the probability of the actual average weight being included within this extended range is also increased. For example, if the estimated range was between 200 and 240 pounds the possibility of accuracy could be increased to the 95% level and if a 190 to 250 pound range had been used a 99% probability level could be used.

In making an estimate it is necessary to decide what probability of being wrong would be acceptable and this in turn will establish the range to be used. The consequence of making a wrong decision will often determine the probability level to be used. If life and death are involved then this probability must be much higher (larger range) than if only economic considerations are at stake.

If no probability of being wrong will be acceptable then the range will have to be extremely large. For example, in our previous problem a range from 0 to 6,000 pounds might be used for the average of the pen of pigs. The actual weight would certainly fall in this range but this type of range makes the estimate of no value in buying hogs.

Relationship of Variance between Population and Sample

If a sample is drawn from a population the sample mean is the best estimate of the population mean. The sample mean would normally not have a value as extreme as the extreme values in the population.

For example:

The population is -- one day's production, in plant A, of bologna as follows:

mean weight: μ = 11 pounds

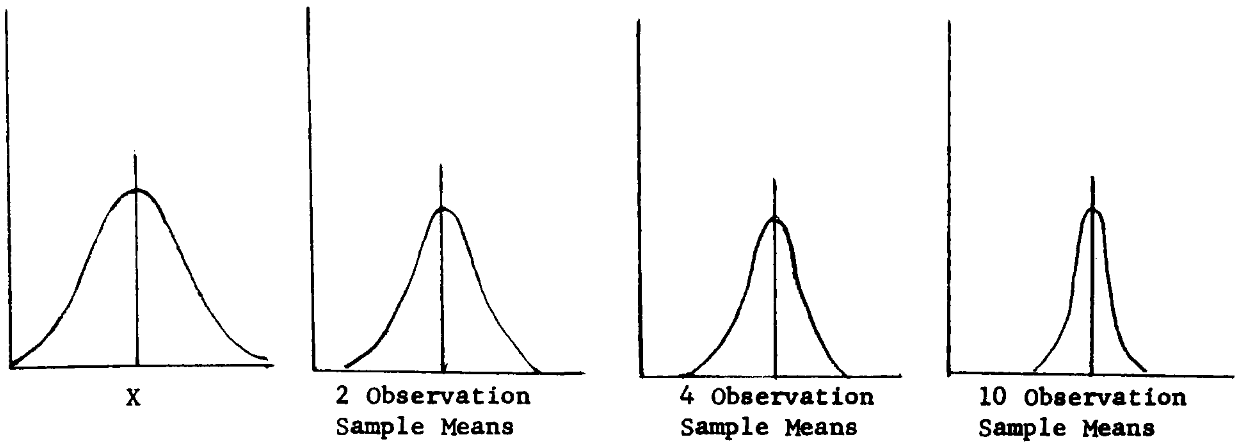
standard deviation: σ = 1 pound

Two casings of bologna were randomly drawn from this production, weighed and returned, and a mean weight was obtained. This operation is repeated several times until there are several sample means from the same population.

The whole sampling procedure is repeated but this time instead of a sample size of two, a sample size of 4 is used. The operation is again repeated and this time a sample size of 10 is used.

If the sample means were now used as raw data, a composite mean of approximately 11 pounds would be expected in all three sample sizes. The mean sample distribution should be clustered more around the population mean value as the sample size becomes larger. The reason for this is there are only a few extreme values in a population and for a mean to have an extreme value all of the individuals drawn for this sample would have to also be extreme values. This is not very probable if the individual samples were randomly selected from the population. It follows that the larger the sample size, the less the probability of the mean having an extreme value.

The distribution of the above population and of sample means (obtained from different size samples) drawn from this population may be illustrated as follows:



Original Population	2 Observation Sample Means	4 Observation Sample Means	10 Observation Sample Means
$\mu = 11$	$\mu_{\bar{X}} = 11$	$\mu_{\bar{X}} = 11$	$\mu_{\bar{X}} = 11$
$\sigma = 1$	$\sigma_{\bar{X}} = .7$	$\sigma_{\bar{X}} = .5$	$\sigma_{\bar{X}} = .3$
$n = 1$	$n = 2$	$n = 4$	$n = 10$

The relationship of the standard deviation of a population to the standard deviation of the means (called the standard error of the mean and represented by the symbol $\sigma_{\bar{X}}$ -also discussed in chapter "Quality Control Charts") drawn from this population is as follows:

Standard error of mean $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

σ = Standard deviation of the population

Calculated from a sample

n = Number of observations in the sample

$$s_{\bar{X}} = \sigma_{\bar{X}} = \sqrt{\frac{\sum (X-\bar{X})^2}{n\,(n-1)}}$$

<u>Student "t" test</u>

The student "t" test allows the testing of a sample to see if it was drawn from a population with a known (or assigned) mean. It is calculated as follows:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

t = "t" test or student test
$\bar{X}$ = sample mean
μ = population mean
$s_{\bar{X}}$ = sample standard error

For example --

10 bologna casings were drawn from a lot -- $n = 10$.

Their average weight was 9.8 pounds -- $\bar{X} = 9.8$

The sample standard error of the mean is .4 -- $s_{\overline{X}} = .4$

Could this sample have been drawn from the previous example in which the population mean was 11 pounds?-- $\mu = 11$

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{9.8 - 11}{.4} = \frac{-1.2}{.4} = -3.0$$

By consulting a "t" table we find that for 9 degrees of freedom (n-1) and a 95% probability level the "t" value would be between -2.26 and +2.26 (table value). The value obtained by the calculation (-3.0) was not within this range so it is assumed that unless a 1 chance in 20 has occurred (95% probability) this sample did not come from a population as described above.

The "t" distribution used in the "t" table is similar to the normal distribution except a few more observations are found in the extreme ranges. This distribution is used in place of the normal distribution when the sample size is small and it approaches the normal distribution when the sample size reaches 30.

Confidence intervals

The confidence interval may be calculated as follows:

Confidence interval (95% probability) $= \overline{X} \pm t_{.05}(s_{\overline{X}})$

$\overline{X}$ = sample mean

$t_{.05}$ = "t" table value at 95% probability; df=n-1

$s_{\overline{X}}$ = sample standard error

For the previous example:

Confidence interval (95% probability) $= \overline{X} \pm t_{.05}\ (s_{\overline{X}})$

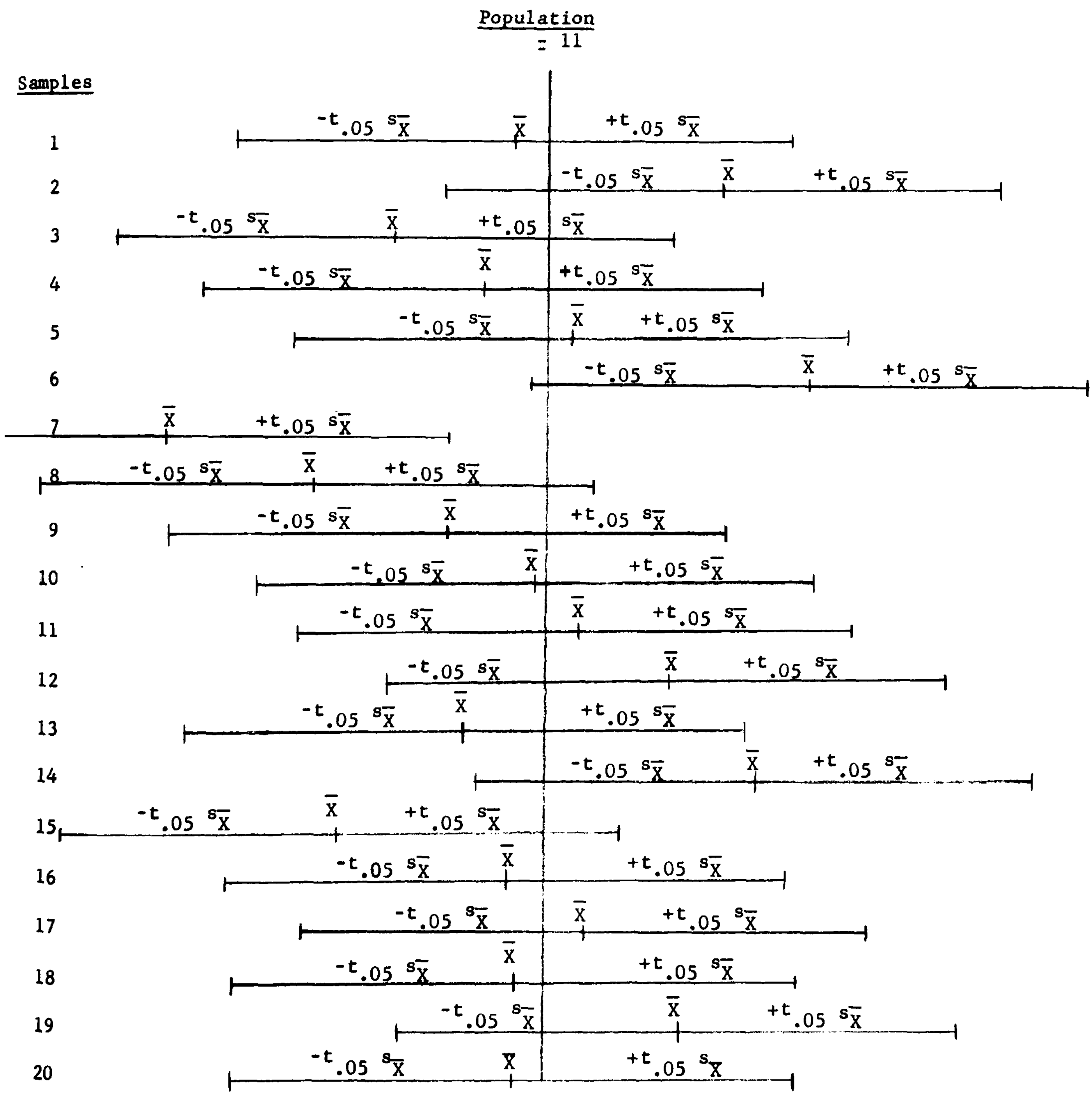
$= 9.8 \pm 2.26\ (.4)$

$= 9.8 \pm .9$

$= 8.9$ to 10.7

This confidence interval indicates that 95% of the time, the population mean from which the sample was drawn would be expected to fall within this range. Since the average value of 11 pounds is not in this range, this agrees with the "t" test.

If the sample had come from the population of -- μ=11, the following distribution of 20 samples drawn from this population would have been expected.



This diagram indicates that the confidence limits of a sample, which is $\bar{X} \pm t_{.05}\, s_{\bar{X}}$, would include the population mean from which it is drawn 19 times out of 20 (95% probability). Sample 7 is the exception in this case. On the other hand if the sample confidence limits does not include the population mean (or an assigned value for it) it is postulated that the sample did not come from that population.

Two Independent Samples

The above procedure can be expanded to indicate if 2 independent samples were drawn from the same population (Are the samples the same?). This procedure is as follows:

Standard error of difference between means:

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\left[\frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}\right]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$ Subscript = sample number

$s_{\bar{X}_1-\bar{X}_2}$ = standard error of difference between means

$$\Sigma x_1^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}$$

Subscript 1 means observations from sample number 1

Subscript 2 means observations from sample number 2

n_1 Number of observations in 1st sample

n_2 Number of observations in 2nd sample

t-test

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

$\bar{X}_1$ = mean of sample number 1

$\bar{X}_2$ = mean of sample number 2

$\mu_1 - \mu_2$ = expected difference between means; zero is often used if the test is to determine if there are any differences between samples.

The "t" obtained by the above test is compared to the table "t" value under the appropriate degrees of freedom (n_1+n_2-2). If the calculated value exceeds the table value (and $\mu_1-\mu_2=0$ was used) then it is postulated that the two sample means came from populations that were different and the samples are significantly different.

Analysis of Variance

Often there is a need to compare more than two means and this is accomplished by the calculation technique known as the analysis of variance. The analysis of variance is based on the following type of model:

Dependent variable = mean + Independent treatments + random error

$Y = \mu$ + Treatment A + ··· Treatment_i + e

Final weight $= \mu$ + Smoking procedure 1 + e
" " 2
" " 3

This data are often referred to on a computer printout sheet as:

Right hand member = μ + Left hand member + e

This analysis is based on two estimates of variance which are:

1. Between group variance:
   a. Variability between the groups
   b. Difference between the means
   c. The larger the difference between means the larger this variance.
2. With group variance:
   a. Variability within groups
   b. Variation of scores within each treatment group
   c. Often called error term

The ratio between these two variances is expressed as:

$$\text{ratio of variance} = \frac{\text{between group variance}}{\text{within group variance}}$$

This ratio is distributed according to the "F" distribution and "F" table values are used to determine if the calculated ratio is large enough to be significant. If the calculated "F" is significant this indicates that there is an <u>overall</u> indication of difference among the means of the groups examined.

An r^2 value can also be calculated in the analysis of variance and it has approximately the same meaning as it had in the correlation area and is calculated as follows:

$$r^2 = \frac{\text{between group sum of squares}}{\text{total sum of squares}}$$

and if r^2 is multiplied by 100 then this is the percentage of the total sums of squares that is accounted for by treatment. This information is important when comparing analysis by two models for the same data. The model which accounts for the largest percentage of the total sum of squares is probably the best model to use to analyze the data.

This analysis is based on the sum of squares of the observations and the total sum of squares is subdivided into two components:

The following example will illustrate the analysis of variance calculation procedure.

Thirty hams were selected that weighed exactly 14 pounds each. These hams were pumped with brine and smoked by three different procedures (labeled group 1, 2 and 3). After smoking, the hams were again weighed and the following results obtained.

Group 1		Group 2		Group 3	
Weight	Weight Squared	Weight	Weight Squared	Weight	Weight Squared
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
13.6	184.96	13.8	190.44	14.2	201.64
14.2	201.64	13.9	193.21	14.1	198.81
14.3	204.49	13.6	184.96	13.8	190.44
14.1	198.81	14.1	198.81	14.4	207.36
13.9	193.21	14.2	201.64	14.3	204.49
13.7	187.69	13.8	190.44	14.1	198.81
14.4	207.36	13.7	187.69	14.0	196.00
14.4	207.36	14.0	196.00	14.1	198.81
13.8	190.44	14.0	196.00	14.2	201.64
14.0	196.00	14.1	198.81	14.2	201.64
$\Sigma X_1=140.4$	$\Sigma X_1^2=1971.96$	$\Sigma X_2=139.2$	$\Sigma X_2^2=1938.00$	$\Sigma X_3=141.4$	$\Sigma X_3^2=1999.64$

$n_1=10$ $\qquad$ $n_2=10$ $\qquad$ $n_3=10$

$\bar{X}_1=14.04$ $\qquad$ $\bar{X}_2=13.92$ $\qquad$ $\bar{X}_3=14.14$

$n_{total} = 30$ $\qquad$ $\bar{X}_{total} = \dfrac{\Sigma X_{total}}{n_{total}} = \dfrac{421.0}{30} = 14.0333$

$$\Sigma X_{total} = \Sigma X_1 + \Sigma X_2 + \Sigma X_3$$

$$= 140.4 + 139.2 + 141.4 = 421.0$$

$$\Sigma X^2{}_{total} = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2$$

$$= 1971.96 + 1938.00 + 1999.64 = 5909.6$$

$$\Sigma x^2{}_{total} = \Sigma X^2{}_{total} - \frac{(\Sigma_X \text{ total})^2}{n_{total}}$$

$$= 5909.6 - \frac{(421)^2}{30} = 5909.6 - \frac{177241}{30}$$

$$= 5909.6 - 5908.0333$$

$$= 1.5667$$

$$\Sigma x^2_{\text{between groups}} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - \frac{(\Sigma X_{total})^2}{n_{total}}$$

$$= \frac{(140.4)^2}{10} + \frac{(139.2)^2}{10} + \frac{(141.4)^2}{10} - \frac{(421.0)^2}{30}$$

$$= \frac{19712.16}{10} + \frac{19376.64}{10} + \frac{19993.96}{10} - \frac{177241.00}{30}$$

$$= 1971.216 + 1937.664 + 1999.396 - 5908.0333$$

$$= .2427$$

$$\Sigma x^2_{\text{within group}} = \left[\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}\right] + \left[\Sigma X_2^2 - \frac{(\Sigma X2)^2}{n_2}\right] + \left[\Sigma X_3^2 - \frac{(\Sigma X_3)^2}{n_3}\right]$$

$$= \left[1,971.96 - \frac{(140.4)^2}{10}\right] + \left[1,938.00 - \frac{(139.2)^2}{10}\right] + \left[1,999.64 - \frac{(141.4)^2}{10}\right]$$

$$= \left[1,971.96 - \frac{19,712.16}{10}\right] + \left[1,938.00 - \frac{19,376.64}{10}\right] + \left[1,999.64 - \frac{19,993.96}{10}\right]$$

$$= \left[1,971.96 - 1,971.216\right] + \left[1,938.00 - 1,937.664\right] + \left[1,999.64 - 1,999.396\right]$$

$$= .744 + .336 + .244$$

$$= 1.324$$

Check on $\Sigma x^2_{\text{within group}}$

$$\Sigma x^2_{\text{within group}} = \Sigma x^2_{\text{total}} - \Sigma x^2_{\text{between group}}$$

$$= 1.5667 - .2427$$

$$= 1.324 \text{ --- agrees with previous results}$$

Check on degrees of freedom

Total df = Between groups df + within groups df

= 2 + 27

= 29 --- agrees with total df

Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom (df)	Mean Square or Variance Estimate	"F" Ratio
Between groups	($\Sigma x^2_{\text{between group}}$) .2427	(Number of groups-1) 3 - 1 = 2	$\left[\frac{\text{Sum of squares}}{\text{df}}\right]$ .2427/2 = .1214	$\left[\frac{\text{Mean square between}}{\text{Mean square within}}\right]$ $\frac{.1214}{.0490} = 2.4776$
Within groups	($\Sigma x^2_{\text{within group}}$) 1.324	(Total Observations - No. of groups) 30 - 3 = 27	$\left[\frac{\text{Sum of squares}}{\text{df}}\right]$ 1.324/27 = .0490	
Total	($\Sigma x^2_{\text{total}}$) 1.5667	(Total Observations - 1) 30 - 1 = 29		

The calculated "F" ratio is 2.48 with 2-27 degrees of freedom.
The table "F" value using the above degrees of freedom (2-27) and the 95% probability level is 3.35.

$$r^2 = \frac{\text{between group sum of squares}}{\text{total sum of squares}} \frac{.2427}{1.5667} = .1549$$

.1549 x 100 = 15.49% of the total sum of squares was accounted for by the group treatments.

The calculated "F" value (2.48) did not exceed the table value (3.35), therefore, the 3 pumping and smoking procedures are not considered to produce finished product weights that are significantly different at the 95% level. This much variation in mean value weights (14.04, 13.92, 14.14) is considered normal for 3 samples (of 10 observations each) being drawn from the same population. No further tests are made when these results are non-significant.

If the calculated value had been significant then the individual treatments (Groups 1, 2, and 3) could have been compared using the "t" or other appropriate tests.

More involved analysis of variance problems with several sub-classifications are possible to analyze but the principles involved are the same as the simple problem just covered.

Sum of Squares Explanation

It is often difficult to visualize what the "Sum of Squares" in the "Analysis of Variance Table" is actually representing. In order to make this calculation a little more enlightening the previous problem will be recalculated using a more difficult calculation technique (because of rounding problems) that is much simpler to visualize.

To calculate the "total sum of squares" each observation is subtracted from the overall mean and this deviation is squared and summed. This represents the total deviation in this set of data.

The "total sum of squares" is recalculated as follows:

Calculation of "total sum of squares" (Each observation is compared to the overall mean)

$\bar{X}_{total} = 14.033333333$

X_1	x_1 $X_1-\bar{X}_{total}$	x_1^2 $(X_1-\bar{X}_{total})^2$	X_2	x_2 $X_2-\bar{X}_{total}$	x_2^2 $(X_2-\bar{X}_{total})^2$	X_3	x_3 $X_2-\bar{X}_{total}$	x_3^2 $(X_3-\bar{X}_{total})^2$
13.6	-.433333333	.187777777	13.8	-.233333333	.054444444	14.2	.166666667	.027777778
14.2	.166666667	.027777778	13.9	-.133333333	.017777778	14.1	.066666667	.004444444
14.3	.266666667	.071111111	13.6	-.433333333	.187777777	13.8	-.233333333	.054444444
14.1	.066666667	.004444444	14.1	.066666667	.004444444	14.4	.366666667	.134444445
13.9	-.133333333	.017777778	14.2	.166666667	.027777778	14.3	.266666667	.071111111
13.7	-.333333333	.111111111	13.8	-.233333333	.054444444	14.1	.066666667	.004444444
14.4	.366666667	.134444445	13.7	-.333333333	.111111111	14.0	-.033333333	.001111111
14.4	.366666667	.134444445	14.0	-.033333333	.001111111	14.1	.066666667	.004444444
13.8	-.233333333	.054444444	14.0	-.033333333	.001111111	14.2	.166666667	.027777778
14.0	-.033333333	.001111111	14.1	.066666667	.004444444	14.2	.166666667	.027777778
	+1.233333335	.744444444		+.300000001	.464444442		+1.333333336	.357777777
	-1.166666665			-1.433333331			- .266666666	
	+ .066666670			-1.133333330			+1.066666670	

$\Sigma X_1-\bar{X}_{total}$ = .066666670 − 1.133333330 + 1.066666670 = .00000001 or 0 except for rounding

Calculation of "total sum of squares" (comparing each value with overall mean)

$\Sigma x^2_{total} = x_1^2 + x_2^2 + x_3^2$ = .744444444 + .464444442 + .357777777 = 1.566666663 or 1.5667

Sum of deviation squared of individual observations from the overall mean; Total deviation if all groups were considered as one population.

The "total sum of squares" representing the total deviation in the data will now be segmented into the deviation "between the groups" (usually caused by treatments) and the deviation "within the groups" (not explained by treatments).

To calculate the "between group sum of squares" the mean for the group containing the observation is subtracted from the overall mean and this deviation is squared and summed. This is repeated for each original observation. This represents the portion of the total deviation accounted for by segregating the data into groups. This in turn can be interpreted as the influence of the treatments on the group means and represents that portion of the total deviation that can be assigned to treatment effects.

The "between group sum of squares" is recalculated as follows:

Calculation of "between groups sum of squares" (If each value was equivalent to its group mean and was compared to the overall mean)

$\bar{X}_{total} = 14.033333333$

X_1 if $=\bar{X}_1$	$X_1-\bar{X}_{total}$	$(X_1-\bar{X}_{total})^2$	X_2 if $=\bar{X}_2$	$X_2-\bar{X}_{total}$	$(X_2-\bar{X}_{total})^2$	X_3 if $=\bar{X}_3$	$X_3-\bar{X}_{total}$	$(X_3-\bar{X}_{total})^2$
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
14.04	.006666667	.000044444	13.92	-.113333333	.012844444	14.14	.106666667	.011377778
	Σ.066666670	Σ.000444440		Σ-1.133333330	Σ.128444440		Σ1.066666670	Σ.113777780

Calculation of "between group sum of squares" (Comparing the group means with the overall means one time for each observation)

$\Sigma(X_i-\bar{X}_{total}) = \Sigma\ (X_1-\bar{X}_{total}) + \Sigma\ (X_2-\bar{X}_{total}) + \Sigma\ (X_3-\bar{X}_{total}) = .066666670-1.133333330+1.066666670=0.00000001$ or 0

Σx^2 between groups $= \Sigma(X_1-\bar{X}_{total})^2 + \Sigma(X_2-\bar{X}_{total})^2 + \Sigma(X_3-\bar{X}_{total})^2$

or $= n_1(X_1-\bar{X}_{total})^2 + n_2(X_2-\bar{X}_{total})^2 + n_3(X_3-\bar{X}_{total})^2$

$= .000444440 + .128444440 + .113777780$

$= .242666660$ or $.2427$

Sum of deviation squared or group means from overall mean once for each observation in each group; portion of total deviation that can be contributed to the group which is usually the result of a treatment.

The "total sum of squares" representing the total deviation has been divided into "between groups sum of squares" and the remainder (within group sum of squares). The remainder which represents the "within group variation" is the portion of the deviation that cannot be accounted for by the groups (by treatments). It is the deviation within treatment groups.

To calculate the "within group" sum of squares the original observation is subtracted from its group mean and this deviation is squared and summed. This is repeated for all observations in all groups. This represents the variation within a group or treatment and is the variation unaccounted for by treatment.

The "within group sum of squares" is recalculated as follows:

Calculation of "within group sum of squares" (Each observation within a group is compared to the group mean)

$$\bar{X}_1 = 14.04; \ \bar{X}_2 = 13.92; \ \bar{X}_3 = 14.14$$

X_1	$X_1-\bar{X}_1$	$(X_1-\bar{X}_1)^2$	X_2	$X_2-\bar{X}_2$	$(X_2-\bar{X}_2)^2$	X_3	$X_3-\bar{X}_3$	$(X_3-\bar{X}_3)^2$
13.6	-.44	.1936	13.8	-.12	.0144	14.2	.06	.0036
14.2	.16	.0256	13.9	-.02	.0004	14.1	-.04	.0016
14.3	.26	.0676	13.6	-.32	.1024	13.8	-.34	.1156
14.1	.06	.0036	14.1	.18	.0324	14.4	.26	.0676
13.9	-.14	.0196	14.2	.28	.0784	14.3	.16	.0256
13.7	-.34	.1156	13.8	-.12	.0144	14.1	-.04	.0016
14.4	.36	.1296	13.7	-.22	.0484	14.0	-.14	.0196
14.4	.36	.1296	14.0	.08	.0064	14.1	-.04	.0016
13.8	-.24	.0576	14.0	.08	.0064	14.2	.06	.0036
14.0	-.04	.0016	14.1	.18	.0324	14.2	.06	.0036
	Σ= 0	Σ =.7440		Σ= 0	Σ =.3360		Σ=0	Σ=.2440

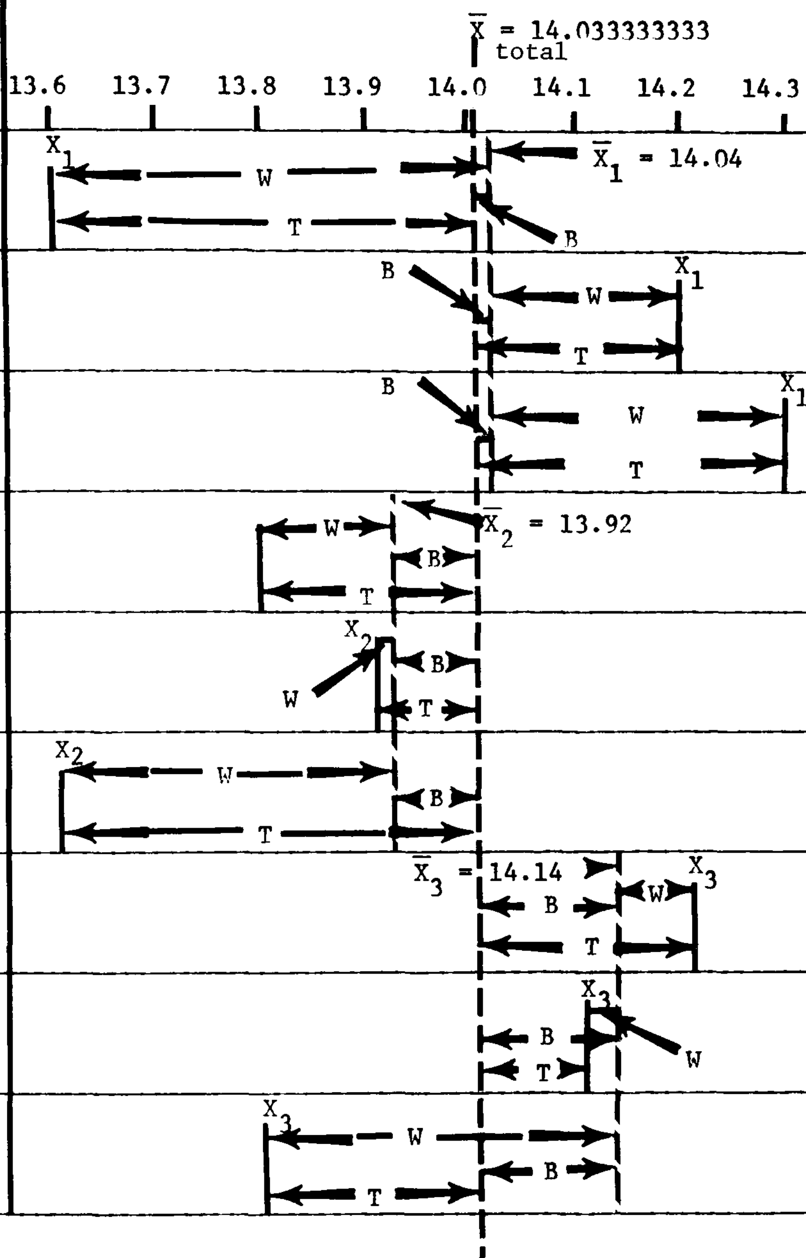
Calculation of "within group sum of squares" (Comparing each observation with the mean of the group it is in)

$$\Sigma x^2 \text{ with in group} = \Sigma(X_1-\bar{X}_1)^2 + \Sigma(X_2-\bar{X}_2)^2 + \Sigma(X_3-\bar{X}_3)^2$$
$$= .7440 + .3360 + .2440$$
$$= 1.324$$

Sum of deviation squared for individual observations from their group mean; portion of total deviation that is within a group; the group or treatment cannot account for this deviation; deviation is caused by something other than group or treatment.

To diagrammatically illustrate the previous calculations (total, between groups, within groups) the first three samples in each group is plotted on a linear scale. The "total deviation", the "between group deviation" and the "within group deviation" is indicated by arrows. Notice that the sum of the "between group deviation" and the "within group deviation" equals the "total deviation". The length of these arrows represents the quantity of deviation in each case. This quantity is then squared and summed for all observations (not just first three), to arrive at the individual totals.

The diagrammatical illustration of these deviations is as follows:



Group	Observation	T Total Deviation $X-\bar{X}_{total}$	B Between Group Deviation $\bar{X}_1-\bar{X}_{total}$	W Within Group Deviation $X_1-\bar{X}_1$	B + W
1	1	-.4333	.0067	-.44	-.4333
1	2	.1667	.0067	.16	.1667
1	3	.2667	.0067	.26	.2667
2	1	-.2333	-.1133	-.12	-.2333
2	2	-.1333	-.1133	-.02	-.1333
2	3	-.4333	-.1133	-.32	-.4333
3	1	.1667	.1067	.06	.1667
3	2	.0667	.1067	-.04	.0667
3	3	-.2333	.1067	-.34	-.2333

Total deviation [deviation between the sample (X) and the total mean ($\bar{X}_{total}$)] is divided into between group deviation [deviation between the treatment mean ($\bar{X}_i$) and the total mean ($\bar{X}_{total}$)] and within group deviation [deviation between the samples (X) and the treatment mean ($\bar{X}_i$)].

Analysis of Variance and Comparison of Means:

Another example of analysis of variance is the following in which tenderizers were evaluated for effectiveness. Twenty steaks were divided into 4 groups of 5 each. Steaks in group one were injected with water (control) and steaks in group 2, 3 and 4 were each injected with a different tenderizer. The steaks were then cooked and sheared (lower shear value greater tenderness).

Group 1 (Control)		Group 2		Group 3		Group 4	
Shear Value	Squared	Shear Value	Squared	Shear Value	Squared	Shear Value	Squared
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
10	100	11	121	8	64	4	16
11	121	11	121	7	49	5	25
12	144	11	121	9	81	4	16
11	121	10	100	10	100	8	64
10	100	9	81	11	121	6	36
$\Sigma X_1=54$	$\Sigma X_1^2=586$	$\Sigma X_2=52$	$\Sigma X_2^2=544$	$\Sigma X_3=45$	$\Sigma X_3^2=415$	$\Sigma X_4=27$	$\Sigma X_4^2=157$

$n_1=5$ $\quad n_2=5$ $\quad n_3=5$ $\quad n_4=5$

$\bar{X}_1=10.8$ $\quad \bar{X}_2=10.4$ $\quad \bar{X}_3=9.0$ $\quad \bar{X}_4=5.4$

$n_{total}=20$

$$\bar{X}_{total} = \frac{\Sigma X_{total}}{n_{total}} = \frac{178}{20} = 8.9$$

$$\Sigma X_{total} = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4$$

$$= 54 + 52 + 45 + 27 = 178$$

$$\Sigma X^2_{total} = \Sigma X_1^{\ 2} + \Sigma X_2^{\ 2} + \Sigma X_3^{\ 2} + \Sigma X_4^{\ 2}$$

$$= 586 + 544 + 415 + 157 = 1702$$

$$\Sigma x^2_{total} = \Sigma X^2_{total} - \frac{(\Sigma X_{total})^2}{n_{total}}$$

$$= 1702 - \frac{(178)^2}{20} = 1702 - \frac{31684}{20}$$

$$= 1702 - 1584.2$$

$$= 117.8$$

$$\Sigma x^2_{\text{between groups}} = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4} - \frac{(\Sigma X_{total})^2}{n_{total}}$$

$$= \frac{(54)^2}{5} + \frac{(52)^2}{5} + \frac{(45)^2}{5} + \frac{(27)^2}{5} - \frac{(178)^2}{20}$$

$$= \frac{2916}{5} + \frac{2704}{5} + \frac{2025}{5} + \frac{729}{5} - \frac{31684}{20}$$

$$= 583.2 + 540.8 + 405.0 + 145.8 - 1584.2$$

$$= 90.6$$

$$\Sigma x^2_{\text{within group}} = \left[\Sigma X_1^{\ 2} - \frac{(\Sigma X_1)^2}{n_1}\right] + \left[\Sigma X_2^{\ 2} - \frac{(\Sigma X_2)^2}{n_2}\right] + \left[\Sigma X_3^{\ 2} - \frac{(\Sigma X_3)^2}{n_3}\right] + \left[\Sigma X_4^{\ 2} - \frac{(\Sigma X_4)^2}{n_4}\right]$$

$$= \left[586 - \frac{(54)^2}{5}\right] + \left[544 - \frac{(52)^2}{5}\right] + \left[415 - \frac{(45)^2}{5}\right] + \left[157 - \frac{(27)^2}{5}\right]$$

$$= \left[586 - \frac{2916}{5}\right] + \left[544 - \frac{2704}{5}\right] + \left[415 - \frac{2025}{5}\right] + \left[157 - \frac{729}{5}\right]$$

$$= \left[586 - 583.2\right] + \left[544 - 540.8\right] + \left[415 - 405\right] + \left[157 - 145.8\right]$$

$$2.8 + 3.2 + 10 + 11.2$$

$$= 27.2$$

Check on $\Sigma x^2_{\text{within group}}$

$$\Sigma x^2_{\text{within group}} = \Sigma x^2_{\text{total}} - \Sigma x^2_{\text{between group}}$$

$$= 117.8 - 90.6$$

$= 27.2$ - - - - - - - - - - agrees with previous results

Check on degrees of freedom

Total df = Between group df + within group df

= 3 + 16

= 19 - - - - - - - - - - agrees with total df

Analysis of Variance Table (tenderizers)

Source of Variation	Sum of Squares	Degrees of Freedom (df)	Mean Square or Variance Estimate	"F" Ratio
Between groups	Σx^2 between group 90.6	Number of groups - 1 = 4 - 1 = 3	$\left[\frac{\text{Sum of Squares}}{\text{df}}\right] =$ 90.6/3 = 30.2	$\frac{\text{Mean Square between}}{\text{Mean Square within}}$ 30.2/1.7 = 17.76
Within groups	Σx^2 within group 27.2	Total observations - number of groups = 20 - 4 = 16	$\left[\frac{\text{Sum of squares}}{\text{df}}\right] =$ 27.2/16 = 1.7	
Total	Σx^2 total 117.8	Total observation - 1 = 20 - 1 = 19		

The calculated "F" ratio is 17.76 with 3 - 16 degrees of freedom.

The table "F" value using 3 - 16 degrees of freedom and the 95% probability table is 3.24 and at 99% probability table values is 5.29. The 95% probability value will be used in the rest of the explanation.

$$r^2 = \frac{\text{between group sum of squares}}{\text{total sum of squares}} = \frac{90.6}{117.8} = .7691$$

.7691 x 100 = 76.91% of total sum of squares is accounted for by the enzyme treatment.

F-distribution

The calculated F is significant if it is equal to or greater than the value shown in table - Upper 5 per cent points (95% level)

Degrees of freedom for greater mean square (ν_1); Degrees of freedom for least mean square (ν_2)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	161·45	199·50	215·71	224·58	230·16	233·99	236·77	238·88	240·54
2	18·513	19·000	19·164	19·247	19·296	19·330	19·353	19·371	19·385
3	10·128	9·5521	9·2766	9·1172	9·0135	8·9406	8·8867	8·8452	8·8123
4	7·7086	6·9443	6·5914	6·3882	6·2561	6·1631	6·0942	6·0410	5·9988
5	6·6079	5·7861	5·4095	5·1922	5·0503	4·9503	4·8759	4·8183	4·7725
6	5·9874	5·1433	4·7571	4·5337	4·3874	4·2839	4·2067	4·1468	4·0990
7	5·5914	4·7374	4·3468	4·1203	3·9715	3·8660	3·7870	3·7257	3·6767
8	5·3177	4·4590	4·0662	3·8379	3·6875	3·5806	3·5005	3·4381	3·3881
9	5·1174	4·2565	3·8625	3·6331	3·4817	3·3738	3·2927	3·2296	3·1789
10	4.9646	4·1028	3·7083	3·4780	3·3258	3·2172	3·1355	3·0717	3·0204
11	4·8443	3·9823	3·5874	3·3567	3·2039	3·0946	3·0123	2·9480	2·8962
12	4.7472	3·8853	3·4903	3·2592	3·1059	2·9961	2·9134	2·8486	2·7964
13	4·6672	3·8056	3·4105	3·1791	3·0254	2·9153	2·8321	2·7669	2·7144
14	4·6001	3·7389	3·3439	3·1122	2·9582	2·8477	2·7642	2·6987	2·6458
15	4·5431	3·6823	3·2874	3·0556	2·9013	2·7905	2·7066	2·6408	2·5876
16	4·4940	3·6337	3·2389	3·0069	2·8524	2·7413	2·6572	2·5911	2·5377
17	4·4513	3·5915	3·1968	2·9647	2·8100	2·6987	2·6143	2·5480	2·4943
18	4·4139	3·5546	3·1599	2·9277	2·7729	2·6613	2·5767	2·5102	2·4563
19	4·3807	3·5219	3·1274	2·8951	2·7401	2·6283	2·5435	2·4768	2·4227
20	4·3512	3·4928	3·0984	2·8661	2·7109	2·5990	2·5140	2·4471	2·3928
21	4·3248	3·4668	3·0725	2·8401	2·6848	2·5727	2·4876	2·4205	2·3660
22	4·3009	3·4434	3·0491	2·8167	2·6613	2·5491	2·4638	2·3965	2·3419
23	4.2793	3·4221	3·0280	2·7955	2·6400	2·5277	2·4422	2·3748	2·3201
24	4·2597	3·4028	3·0088	2·7763	2·6207	2·5082	2·4226	2·3551	2·3002
25	4·2417	3·3852	2·9912	2·7587	2·6030	2·4904	2·4047	2·3371	2·2821
26	4·2252	3·3690	2·9752	2·7426	2·5868	2·4741	2·3883	2·3205	2·2655
27	4·2100	3·3541	2·9604	2·7278	2·5719	2·4591	2·3732	2·3053	2·2501
28	4·1960	3·3404	2·9467	2·7141	2·5581	2·4453	2·3593	2·2913	2·2360
29	4·1830	3·3277	2·9340	2·7014	2·5454	2·4324	2·3463	2·2783	2·2229
30	4·1709	3·3158	2·9223	2·6896	2·5336	2·4205	2·3343	2·2662	2·2107
40	4·0847	3·2317	2·8387	2·6060	2·4495	2·3359	2·2490	2·1802	2·1240
60	4·0012	3·1504	2·7581	2·5252	2·3683	2·2541	2·1665	2·0970	2·0401
120	3·9201	3·0718	2·6802	2·4472	2·2899	2·1750	2·0868	2·0164	1·9588
∞	3·8415	2·9957	2·6049	2·3719	2·2141	2·0986	2·0096	1·9384	1·8799

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	241·88	243·91	245·95	248·01	249·05	250·10	251·14	252·20	253·25	254·31
2	19·396	19·413	19·429	19·446	19·454	19·462	19·471	19·479	19·487	19·496
3	8·7855	8·7446	8·7029	8·6602	8·6385	8·6166	8·5944	8·5720	8·5494	8·5264
4	5·9644	5·9117	5·8578	5·8025	5·7744	5·7459	5·7170	5·6877	5·6581	5·6281
5	4·7351	4·6777	4·6188	4·5581	4·5272	4·4957	4·4638	4·4314	4·3985	4·3650
6	4·0600	3·9999	3·9381	3·8742	3·8415	3·8082	3·7743	3·7398	3·7047	3·6689
7	3·6365	3·5747	3·5107	3·4445	3·4105	3·3758	3·3404	3·3043	3·2674	3·2298
8	3·3472	3·2839	3·2184	3·1503	3·1152	3·0794	3·0428	3·0053	2·9669	2·9276
9	3·1373	3·0729	3·0061	2·9365	2·9005	2·8637	2·8259	2·7872	2·7475	2·7067
10	2·9782	2·9130	2·8450	2·7740	2·7372	2·6996	2·6609	2·6211	2·5801	2·5379
11	2·8536	2·7876	2·7186	2·6464	2·6090	2·5705	2·5309	2·4901	2·4480	2·4045
12	2·7534	2·6866	2·6169	2·5436	2·5055	2·4663	2·4259	2·3842	2·3410	2·2962
13	2·6710	2·6037	2·5331	2·4589	2·4202	2·3803	2·3392	2·2966	2·2524	2·2064
14	2·6022	2·5342	2·4630	2·3879	2·3487	2·3082	2·2664	2·2229	2·1778	2·1307
15	2·5437	2·4753	2·4034	2·3275	2·2878	2·2468	2·2043	2·1601	2·1141	2·0658
16	2·4935	2·4247	2·3522	2·2756	2·2354	2·1938	2·1507	2·1058	2·0589	2·0096
17	2·4499	2·3807	2·3077	2·2304	2·1898	2·1477	2·1040	2·0584	2·0107	1·9604
18	2·4117	2·3421	2·2686	2·1906	2·1497	2·1071	2·0629	2·0166	1·9681	1·9168
19	2·3779	2·3080	2·2341	2·1555	2·1141	2·0712	2·0264	1·9795	1·9302	1·8780
20	2·3479	2·2776	2·2033	2·1242	2·0825	2·0391	1·9938	1·9464	1·8963	1·8432
21	2·3210	2·2504	2·1757	2·0960	2·0540	2·0102	1·9645	1·9165	1·8657	1·8117
22	2·2967	2·2258	2·1508	2·0707	2·0283	1·9842	1·9380	1·8894	1·8380	1·7831
23	2·2747	2·2036	2·1282	2·0476	2·0050	1·9605	1·9139	1·8648	1·8128	1·7570
24	2·2547	2·1834	2·1077	2·0267	1·9838	1·9390	1·8920	1·8424	1·7896	1·7330
25	2·2365	2·1649	2·0889	2·0075	1·9643	1·9192	1·8718	1·8217	1·7684	1·7110
26	2·2197	2·1479	2·0716	1·9898	1·9464	1·9010	1·8533	1·8027	1·7488	1·6906
27	2·2043	2·1323	2·0558	1·9736	1·9299	1·8842	1·8361	1·7851	1·7306	1·6717
28	2·1900	2·1179	2·0411	1·9586	1·9147	1·8687	1·8203	1·7689	1·7138	1·6541
29	2·1768	2·1045	2·0275	1·9446	1·9005	1·8543	1·8055	1·7537	1·6981	1·6376
30	2·1646	2·0921	2·0148	1·9317	1·8874	1·8409	1·7918	1·7396	1·6835	1·6223
40	2·0772	2·0035	1·9245	1·8389	1·7929	1·7444	1·6928	1·6373	1·5766	1·5089
60	1·9926	1·9174	1·8364	1·7480	1·7001	1·6491	1·5943	1·5343	1·4673	1·3893
120	1·9105	1·8337	1·7505	1·6587	1·6084	1·5543	1·4952	1·4290	1·3519	1·2539
∞	1·8307	1·7522	1·6664	1·5705	1·5173	1·4591	1·3940	1·3180	1·2214	1·0000

Source: Biometrika Trustees Biometrika Tables for Statisticians, Vol. II, 1972. Ed. E. S. Pearson & H. D. Hartley

F-distribution

The calculated F is significant if it is equal to or greater than the value shown in table - Upper 1 per cent points (99% level)

Degrees of freedom for lesser mean square (ν_2) / Degrees of freedom for greater mean square (ν_1)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	4052·2	4999·5	5403·4	5624·6	5763·6	5859·0	5928·4	5981·1	6022·5
2	98·503	99·000	99·166	99·249	99·299	99·333	99·356	99·374	99·388
3	34·116	30·817	29·457	28·710	28·237	27·911	27·672	27·489	27·345
4	21·198	18·000	16·694	15·977	15·522	15·207	14·976	14·799	14·659
5	16·258	13·274	12·060	11·392	10·967	10·672	10·456	10·289	10·158
6	13·745	10·925	9·7795	9·1483	8·7459	8·4661	8·2600	8·1017	7·9761
7	12·246	9·5466	8·4513	7·8466	7·4604	7·1914	6·9928	6·8400	6·7188
8	11·259	8·6491	7·5910	7·0061	6·6318	6·3707	6·1776	6·0289	5·9106
9	10·561	8·0215	6·9919	6·4221	6·0569	5·8018	5·6129	5·4671	5·3511
10	10·044	7·5594	6·5523	5·9943	5·6363	5·3858	5·2001	5·0567	4·9424
11	9·6460	7·2057	6·2167	5·6683	5·3160	5·0692	4·8861	4·7445	4·6315
12	9·3302	6·9266	5·9525	5·4120	5·0643	4·8206	4·6395	4·4994	4·3875
13	9·0738	6·7010	5·7394	5·2053	4·8616	4·6204	4·4410	4·3021	4·1911
14	8·8616	6·5149	5·5639	5·0354	4·6950	4·4558	4·2779	4·1399	4·0297
15	8·6831	6·3589	5·4170	4·8932	4·5556	4·3183	4·1415	4·0045	3·8948
16	8·5310	6·2262	5·2922	4·7726	4·4374	4·2016	4·0259	3·8896	3·7804
17	8·3997	6·1121	5·1850	4·6690	4·3359	4·1015	3·9267	3·7910	3·6822
18	8·2854	6·0129	5·0919	4·5790	4·2479	4·0146	3·8406	3·7054	3·5971
19	8·1849	5·9259	5·0103	4·5003	4·1708	3·9386	3·7653	3·6305	3·5225
20	8·0960	5·8489	4·9382	4·4307	4·1027	3·8714	3·6987	3·5644	3·4567
21	8·0166	5·7804	4·8740	4·3688	4·0421	3·8117	3·6396	3·5056	3·3981
22	7·9454	5·7190	4·8166	4·3134	3·9880	3·7583	3·5867	3·4530	3·3458
23	7·8811	5·6637	4·7649	4·2636	3·9392	3·7102	3·5390	3·4057	3·2986
24	7·8229	5·6136	4·7181	4·2184	3·8951	3·6667	3·4959	3·3629	3·2560
25	7·7698	5·5680	4·6755	4·1774	3·8550	3·6272	3·4568	3·3239	3·2172
26	7·7213	5·5263	4·6366	4·1400	3·8183	3·5911	3·4210	3·2884	3·1818
27	7·6767	5·4881	4·6009	4·1056	3·7848	3·5580	3·3882	3·2558	3·1494
28	7·6356	5·4529	4·5681	4·0740	3·7539	3·5276	3·3581	3·2259	3·1195
29	7·5977	5·4204	4·5378	4·0449	3·7254	3·4995	3·3303	3·1982	3·0920
30	7·5625	5·3903	4·5097	4·0179	3·6990	3·4735	3·3045	3·1726	3·0665
40	7·3141	5·1785	4·3126	3·8283	3·5138	3·2910	3·1238	2·9930	2·8876
60	7·0771	4·9774	4·1259	3·6490	3·3389	3·1187	2·9530	2·8233	2·7185
120	6·8509	4·7865	3·9491	3·4795	3·1735	2·9559	2·7918	2·6629	2·5586
∞	6·6349	4·6052	3·7816	3·3192	3·0173	2·8020	2·6393	2·5113	2·4073

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	6055·8	6106·3	6157·3	6208·7	6234·6	6260·6	6286·8	6313·0	6339·4	6365·9
2	99·399	99·416	99·433	99·449	99·458	99·466	99·474	99·482	99·491	99·499
3	27·229	27·052	26·872	26·690	26·598	26·505	26·411	26·316	26·221	26·125
4	14·546	14·374	14·198	14·020	13·929	13·838	13·745	13·652	13·558	13·463
5	10·051	9·8883	9·7222	9·5526	9·4665	9·3793	9·2912	9·2020	9·1118	9·0204
6	7·8741	7·7183	7·5590	7·3958	7·3127	7·2285	7·1432	7·0567	6·9690	6·8800
7	6·6201	6·4691	6·3143	6·1554	6·0743	5·9920	5·9084	5·8236	5·7373	5·6495
8	5·8143	5·6667	5·5151	5·3591	5·2793	5·1981	5·1156	5·0316	4·9461	4·8588
9	5·2565	5·1114	4·9621	4·8080	4·7290	4·6486	4·5666	4·4831	4·3978	4·3105
10	4·8491	4·7059	4·5581	4·4054	4·3269	4·2469	4·1653	4·0819	3·9965	3·9090
11	4·5393	4·3974	4·2509	4·0990	4·0209	3·9411	3·8596	3·7761	3·6904	3·6024
12	4·2961	4·1553	4·0096	3·8584	3·7805	3·7008	3·6192	3·5355	3·4494	3·3608
13	4·1003	3·9603	3·8154	3·6646	3·5868	3·5070	3·4253	3·3413	3·2548	3·1654
14	3·9394	3·8001	3·6557	3·5052	3·4274	3·3476	3·2656	3·1813	3·0942	3·0040
15	3·8049	3·6662	3·5222	3·3719	3·2940	3·2141	3·1319	3·0471	2·9595	2·8684
16	3·6909	3·5527	3·4089	3·2587	3·1808	3·1007	3·0182	2·9330	2·8447	2·7528
17	3·5931	3·4552	3·3117	3·1615	3·0835	3·0032	2·9205	2·8348	2·7459	2·6530
18	3·5082	3·3706	3·2273	3·0771	2·9990	2·9185	2·8354	2·7493	2·6597	2·5660
19	3·4338	3·2965	3·1533	3·0031	2·9249	2·8442	2·7608	2·6742	2·5839	2·4893
20	3·3682	3·2311	3·0880	2·9377	2·8594	2·7785	2·6947	2·6077	2·5168	2·4212
21	3·3098	3·1730	3·0300	2·8796	2·8010	2·7200	2·6359	2·5484	2·4568	2·3603
22	3·2576	3·1209	2·9779	2·8274	2·7488	2·6675	2·5831	2·4951	2·4029	2·3055
23	3·2106	3·0740	2·9311	2·7805	2·7017	2·6202	2·5355	2·4471	2·3542	2·2558
24	3·1681	3·0316	2·8887	2·7380	2·6591	2·5773	2·4923	2·4035	2·3100	2·2107
25	3·1294	2·9931	2·8502	2·6993	2·6203	2·5383	2·4530	2·3637	2·2696	2·1694
26	3·0941	2·9578	2·8150	2·6640	2·5848	2·5026	2·4170	2·3273	2·2325	2·1315
27	3·0618	2·9256	2·7827	2·6316	2·5522	2·4699	2·3840	2·2938	2·1985	2·0965
28	3·0320	2·8959	2·7530	2·6017	2·5223	2·4397	2·3535	2·2629	2·1670	2·0642
29	3·0045	2·8685	2·7256	2·5742	2·4946	2·4118	2·3253	2·2344	2·1379	2·0342
30	2·9791	2·8431	2·7002	2·5487	2·4689	2·3860	2·2992	2·2079	2·1108	2·0062
40	2·8005	2·6648	2·5216	2·3689	2·2880	2·2034	2·1142	2·0194	1·9172	1·8047
60	2·6318	2·4961	2·3523	2·1978	2·1154	2·0285	1·9360	1·8363	1·7263	1·6006
120	2·4721	2·3363	2·1915	2·0346	1·9500	1·8600	1·7628	1·6557	1·5330	1·3805
∞	2·3209	2·1847	2·0385	1·8783	1·7908	1·6964	1·5923	1·4730	1·3246	1·0000

Source: Biometrika Trustees Biometrika Tables for Statisticians, Vol. II, 1972. Ed. E. S. Pearson & H. D. Hartley.

F-distribution

The calculated F is significant if it is equal to or greater than the value shown in table - Upper 0.5 percent points (99.5% level)

Degrees of freedom for greater mean square (ν_1); Degrees of freedom for lesser mean square (ν_2)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	198·50	199·00	199·17	199·25	199·30	199·33	199·36	199·37	199·39
3	55·552	49·799	47·467	46·195	45·392	44·838	44·434	44·126	43·882
4	31·333	26·284	24·259	23·155	22·456	21·975	21·622	21·352	21·139
5	22·785	18·314	16·530	15·556	14·940	14·513	14·200	13·961	13·772
6	18·635	14·544	12·917	12·028	11·464	11·073	10·786	10·566	10·391
7	16·236	12·404	10·882	10·050	9·5221	9·1553	8·8854	8·6781	8·5138
8	14·688	11.042	9·5965	8·8051	8·3018	7·9520	7·6941	7·4959	7·3386
9	13·614	10·107	8·7171	7·9559	7·4712	7·1339	6·8849	6·6933	6·5411
10	12·826	9·4270	8·0807	7·3428	6·8724	6·5446	6·3025	6·1159	5·9676
11	12·226	8·9122	7·6004	6·8809	6·4217	6·1016	5·8648	5·6821	5·5368
12	11·754	8·5096	7·2258	6·5211	6·0711	5·7570	5·5245	5·3451	5·2021
13	11·374	8·1865	6·9258	6·2335	5·7910	5·4819	5·2529	5·0761	4·9351
14	11·060	7·9216	6·6804	5·9984	5·5623	5·2574	5·0313	4·8566	4·7173
15	10·798	7·7008	6·4760	5·8029	5·3721	5·0708	4·8473	4·6744	4·5364
16	10·575	7·5138	6·3034	5·6378	5·2117	4·9134	4·6920	4·5207	4·3838
17	10·384	7·3536	6·1556	5·4967	5·0746	4·7789	4·5594	4·3894	4·2535
18	10·218	7·2148	6·0278	5·3746	4·9560	4·6627	4·4448	4·2759	4·1410
19	10·073	7·0935	5·9161	5·2681	4·8526	4·5614	4·3448	4·1770	4·0428
20	9·9439	6·9865	5·8177	5·1743	4·7616	4·4721	4·2569	4·0900	3·9564
21	9·8295	6·8914	5·7304	5·0911	4·6809	4·3931	4·1789	4·0128	3·8799
22	9·7271	6·8064	5·6524	5·0168	4·6088	4·3225	4·1094	3·9440	3·8116
23	9·6348	6·7300	5·5823	4·9500	4·5441	4·2591	4·0469	3·8822	3·7502
24	9·5513	6·6609	5·5190	4·8898	4·4857	4·2019	3·9905	3·8264	3·6949
25	9·4753	6·5982	5·4615	4·8351	4·4327	4·1500	3·9394	3·7758	3·6447
26	9·4059	6·5409	5·4091	4·7852	4·3844	4·1027	3·8928	3·7297	3·5989
27	9·3423	6·4885	5·3611	4·7396	4·3402	4·0594	3·8501	3·6875	3·5571
28	9·2838	6·4403	5·3170	4·6977	4·2996	4·0197	3·8110	3·6487	3·5186
29	9·2297	6·3958	5·2764	4·6591	4·2622	3·9831	3·7749	3·6131	3·4832
30	9·1797	6·3547	5·2388	4·6234	4·2276	3·9492	3·7416	3·5801	3·4505
40	8·8279	6·0664	4·9758	4·3738	3·9860	3·7129	3·5088	3·3498	3·2220
60	8·4946	5·7950	4·7290	4·1399	3·7599	3·4918	3·2911	3·1344	3·0083
120	8·1788	5·5393	4·4972	3·9207	3·5482	3·2849	3·0874	2·9330	2·8083
∞	7·8794	5·2983	4·2794	3·7151	3·3499	3·0913	2·8968	2·7444	2·6210

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	24224	24426	24630	24836	24940	25044	25148	25253	25359	25464
2	199·40	199·42	199·43	199·45	199·46	199·47	199·47	199·48	199·49	199·50
3	43·686	43·387	43·085	42·778	42·622	42·466	42·308	42·149	41·989	41·828
4	20·967	20·705	20·438	20·167	20·030	19·892	19·752	19·611	19·468	19·325
5	13·618	13·384	13·146	12·903	12·780	12·656	12·530	12·402	12·274	12·144
6	10·250	10·034	9·8140	9·5888	9·4742	9·3582	9·2408	9·1219	9·0015	8·8793
7	8·3803	8·1764	7·9678	7·7540	7·6450	7·5345	7·4224	7·3088	7·1933	7·0760
8	7·2106	7·0149	6·8143	6·6082	6·5029	6·3961	6·2875	6·1772	6·0649	5·9506
9	6·4172	6·2274	6·0325	5·8318	5·7292	5·6248	5·5186	5·4104	5·3001	5·1875
10	5·8467	5·6613	5·4707	5·2740	5·1732	5·0706	4·9659	4·8592	4·7501	4·6385
11	5·4183	5·2363	5·0489	4·8552	4·7557	4·6543	4·5508	4·4450	4·3367	4·2255
12	5·0855	4·9062	4·7213	4·5299	4·4314	4·3309	4·2282	4·1229	4·0149	3·9039
13	4·8199	4·6429	4·4600	4·2703	4·1726	4·0727	3·9704	3·8655	3·7577	3·6465
14	4·6034	4·4281	4·2468	4·0585	3·9614	3·8619	3·7600	3·6552	3·5473	3·4359
15	4·4235	4·2497	4·0698	3·8826	3·7859	3·6867	3·5850	3·4803	3·3722	3·2602
16	4·2719	4·0994	3·9205	3·7342	3·6378	3·5389	3·4372	3·3324	3·2240	3·1115
17	4·1424	3·9709	3·7929	3·6073	3·5112	3·4124	3·3108	3·2058	3·0971	2·9839
18	4·0305	3·8599	3·6827	3·4977	3·4017	3·3030	3·2014	3·0962	2·9871	2·8732
19	3·9329	3·7631	3·5866	3·4020	3·3062	3·2075	3·1058	3·0004	2·8908	2·7762
20	3·8470	3·6779	3·5020	3·3178	3·2220	3·1234	3·0215	2·9159	2·8058	2·6904
21	3·7709	3·6024	3·4270	3·2431	3·1474	3·0488	2·9467	2·8408	2·7302	2·6140
22	3·7030	3·5350	3·3600	3·1764	3·0807	2·9821	2·8799	2·7736	2·6625	2·5455
23	3·6420	3·4745	3·2999	3·1165	3·0208	2·9221	2·8197	2·7132	2·6015	2·4837
24	3·5870	3·4199	3·2456	3·0624	2·9667	2·8679	2·7654	2·6585	2·5463	2·4276
25	3·5370	3·3704	3·1963	3·0133	2·9176	2·8187	2·7160	2·6088	2·4961	2·3765
26	3·4916	3·3252	3·1515	2·9685	2·8728	2·7738	2·6709	2·5633	2·4501	2·3297
27	3·4499	3·2839	3·1104	2·9275	2·8318	2·7327	2·6296	2·5217	2·4079	2·2867
28	3·4117	3·2460	3·0727	2·8899	2·7941	2·6949	2·5916	2·4834	2·3690	2·2470
29	3·3765	3·2110	3·0379	2·8551	2·7594	2·6600	2·5565	2·4479	2·3331	2·2102
30	3·3440	3·1787	3·0057	2·8230	2·7272	2·6278	2·5241	2·4151	2·2998	2·1760
40	3·1167	2·9531	2·7811	2·5984	2·5020	2·4015	2·2958	2·1838	2·0636	1·9318
60	2·9042	2·7419	2·5705	2·3872	2·2898	2·1874	2·0789	1·9622	1·8341	1·6885
120	2·7052	2·5439	2·3727	2·1881	2·0890	1·9840	1·8709	1·7469	1·6055	1·4311
∞	2·5188	2·3583	2·1868	1·9998	1·8983	1·7891	1·6691	1·5325	1·3637	1·0000

Source: Biometrika Trustees Biometrika Tables for Statisticians, Vol. II, 1972. Ed. E. S. Pearson & H. D. Hartley.

The calculated "F" value (17.76) did exceed the table value (3.24), therefore the 4 treatments did produce a significant (95% level) change in tenderness shear values. This much variation in shear values (10.8, 10.4, 9, 5.4) would not be expected by chance alone if the treatments (of 5 observations each) had no effect.

The Analysis of Variance indicates there are significant differences between some of the means. However this test does not indicate which of the means are significantly different from each other. The number of possible comparisons are indicated by the following formula:

$$\text{Number of comparisons} = \frac{(\text{Number of groups})\ (\text{Number of groups-1})}{2}$$

$$= \frac{(4)\ (4-1)}{2} = \frac{(4)\ (3)}{2} = \frac{12}{2}$$

$$= 6$$

These 6 comparisons would be

Group 1 ($\bar{X}$=10.8) with group 2 ($\bar{X}$=10.4)

Group 1 ($\bar{X}$=10.8) with group 3 ($\bar{X}$=9.0)

Group 1 ($\bar{X}$=10.8) with group 4 ($\bar{X}$=5.4)

Group 2 ($\bar{X}$=10.4) with group 3 ($\bar{X}$=9.0)

Group 2 ($\bar{X}$=10.4) with group 4 ($\bar{X}$=5.4)

Group 3 ($\bar{X}$=9.0) with group 4 ($\bar{X}$=5.4)

Several techniques have been devised to determine which of these 6 comparisons are significantly different and each of these techniques are supported by a number of statisticians. The example illustrated here is known as the Duncan Multiple Range Test.

Duncan Multiple Range Test:

This analysis is used only after a treatment has been proven to be significant by the analysis of variance test.

The next step is to line the means (least square means can be used) up in order of magnitude; horizontally high to low (descending order) goes from left to right and virtually low to high (ascended order) from top to bottom as shown in the following example:

Means or least square means

descending order →

Ascending order ↓	10.8	10.4	9.0	5.4
5.4				
9.0				
10.4				
10.8				

The size of the means will not always be in the same order as group numbers as is the case in this example. Use size of means regardless of group number to construct the previous table.

The next step is to substract the vertical value from the horizontal value and place this where the horizontal and vertical axis intersect as shown in the following example:

Means or least square means

	10.8	10.4	9.0	5.4
5.4	5.4	5.0	3.6	-
9.0	1.8	1.4	-	
10.4	0.4	-		
10.8	-			

Subtraction of mean differences

These subtractions compare the means (indicate their differences) of the 6 possible comparisons shown previously.

The next step is to determine the degrees of freedom to use in the Multiple Range Test Table. The first value used is the degrees of freedom of the "within group" (sometimes called error term or remainder) found in the analysis of variance table and in this example it is 16. The next values needed are determined by the counting distance the means are separated from each other and would be as follows:

Means or least square means

	10.8	10.4	9.0	5.4
5.4	4	3	2	-
9.0	3	2	-	
10.4	2	-		
10.8	-			

Numerical distance that separates means

Example Mean 10.8 is 4 means from mean 5.4.

These 2 df values (16 and 4 or 3 or 2) are used in the Multiple Range Test Table to determine a factor than is multiplied by the standard error to determine the distance required between two means to be significant.

$$\text{Standard Error} = S_{\overline{X}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{\text{Mean Square within groups or error term or remainder}}{\text{Number of samples in a group}}}$$

$$= \sqrt{\frac{1.7}{5}} = \sqrt{.34}$$

$$= .58$$

If there are unequal numbers of observations per treatment then this formula becomes:

$$S_{\overline{X}} = \sqrt{\frac{s^2}{\frac{2(n_{\text{group X}})(n_{\text{group Y}})}{(n_{\text{group X}}) + (n_{\text{group Y}})}}}$$

For example if there had been 6 shear values in group 1 and 3 shear values in group 4, the formula would have been:

$$S_{\overline{X}} = \sqrt{\frac{1.7}{\frac{2(6)(3)}{6 + 3}}}$$

and the $S_{\overline{X}}$ value would have to be recalculated for each mean comparison. However for this example we will continue with an equal number in each group.

Distance required between means to be significant = Multiple Range Test Table Value X Standard Error

Distance Means are Separated from Each Other 5% Level
New* Multiple Range Test (Duncan)

Degrees of Freedom of Within Group (error term) Found in Analysis of Variance Table

n_2 \ p	2	3	4	5	6	7	8	9	10	12	14	16	18	20	50	100
1	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0
2	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09
3	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50
4	3.93	4.01	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02
5	3.64	3.74	3.79	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83
6	3.46	3.58	3.64	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68
7	3.35	3.47	3.54	3.58	3.60	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.61
8	3.26	3.39	3.47	3.52	3.55	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
9	3.20	3.34	3.41	3.47	3.50	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52
10	3.15	3.30	3.37	3.43	3.46	3.47	3.47	3.47	3.47	3.47	3.47	3.47	3.47	3.48	3.48	3.48
11	3.11	3.27	3.35	3.39	3.43	3.44	3.45	3.46	3.46	3.46	3.46	3.46	3.47	3.48	3.48	3.48
12	3.08	3.23	3.33	3.36	3.40	3.42	3.44	3.44	3.46	3.46	3.46	3.46	3.47	3.48	3.48	3.48
13	3.06	3.21	3.30	3.35	3.38	3.41	3.42	3.44	3.45	3.45	3.46	3.46	3.47	3.47	3.47	3.47
14	3.03	3.18	3.27	3.33	3.37	3.39	3.41	3.42	3.44	3.45	3.46	3.46	3.47	3.47	3.47	3.47
15	3.01	3.16	3.25	3.31	3.36	3.38	3.40	3.42	3.43	3.44	3.45	3.46	3.47	3.47	3.47	3.47
16	3.00	3.15	3.23	3.30	3.34	3.37	3.39	3.41	3.43	3.44	3.45	3.46	3.47	3.47	3.47	3.47
17	2.98	3.13	3.22	3.28	3.33	3.36	3.38	3.40	3.42	3.44	3.45	3.46	3.47	3.47	3.47	3.47
18	2.97	3.12	3.21	3.27	3.32	3.35	3.37	3.39	3.41	3.43	3.45	3.46	3.47	3.47	3.47	3.47
19	2.96	3.11	3.19	3.26	3.31	3.35	3.37	3.39	3.41	3.43	3.44	3.46	3.47	3.47	3.47	3.47
20	2.95	3.10	3.18	3.25	3.30	3.34	3.36	3.38	3.40	3.43	3.44	3.46	3.46	3.47	3.47	3.47
22	2.93	3.08	3.17	3.24	3.29	3.32	3.35	3.37	3.39	3.42	3.44	3.45	3.46	3.47	3.47	3.47
24	2.92	3.07	3.15	3.22	3.28	3.31	3.34	3.37	3.38	3.41	3.44	3.45	3.46	3.47	3.47	3.47
26	2.91	3.06	3.14	3.21	3.27	3.30	3.34	3.36	3.38	3.41	3.43	3.45	3.46	3.47	3.47	3.47
28	2.90	3.04	3.13	3.20	3.26	3.30	3.33	3.35	3.37	3.40	3.43	3.45	3.46	3.47	3.47	3.47
30	2.89	3.04	3.12	3.20	3.25	3.29	3.32	3.35	3.37	3.40	3.43	3.44	3.46	3.47	3.47	3.47
40	2.86	3.01	3.10	3.17	3.22	3.27	3.30	3.33	3.35	3.39	3.42	3.44	3.46	3.47	3.47	3.47
60	2.83	2.98	3.08	3.14	3.20	3.24	3.28	3.31	3.33	3.37	3.40	3.43	3.45	3.47	3.48	3.48
100	2.80	2.95	3.05	3.12	3.18	3.22	3.26	3.20	3.32	3.36	3.40	3.42	3.45	3.47	3.53	3.53
∞	2.77	2.92	3.02	3.09	3.15	3.19	3.23	3.20	3.20	3.34	3.38	3.41	3.44	3.47	3.61	3.67

*Using special protection levels based on degrees of freedom.

Source; Duncan, D. B., Biometrics, Vol II, 1955

Distance Means are Separated from Each Other 1% Level
New* Multiple Range Test (Duncan)

Degrees of Freedom of Within Group (error term) Found in Analysis of Variance Table

n_2 \ p	2	3	4	5	6	7	8	9	10	12	14	16	18	20	50	100
1	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
2	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0
3	8.26	8.5	8.6	8.7	8.8	8.9	8.9	9.0	9.0	9.0	9.1	9.2	9.3	9.3	9.3	9.3
4	6.51	6.8	6.9	7.0	7.1	7.1	7.2	7.2	7.3	7.3	7.4	7.4	7.5	7.5	7.5	7.5
5	5.70	5.96	6.11	6.18	6.26	6.33	6.40	6.44	6.5	6.6	6.6	6.7	6.7	6.8	6.8	6.8
6	5.24	5.51	5.65	5.73	5.81	5.88	5.95	6.00	6.0	6.1	6.2	6.2	6.3	6.3	6.3	6.3
7	4.95	5.22	5.37	5.45	5.53	5.61	5.69	5.73	5.8	5.8	5.9	5.9	6.0	6.0	6.0	6.0
8	4.74	5.00	5.14	5.23	5.32	5.40	5.47	5.51	5.5	5.6	5.7	5.7	5.8	5.8	5.8	5.8
9	4.60	4.86	4.99	5.08	5.17	5.25	5.32	5.36	5.4	5.5	5.5	5.6	5.7	5.7	5.7	5.7
10	4.48	4.73	4.88	4.96	5.06	5.13	5.20	5.24	5.28	5.36	5.42	5.48	5.54	5.55	5.55	5.55
11	4.39	4.63	4.77	4.86	4.94	5.01	5.06	5.12	5.15	5.24	5.28	5.34	5.38	5.39	5.39	5.39
12	4.32	4.55	4.66	4.76	4.84	4.92	4.96	5.02	5.07	5.13	5.17	5.22	5.24	5.26	5.26	5.26
13	4.26	4.48	4.62	4.69	4.74	4.84	4.88	4.94	4.98	5.04	5.08	5.13	5.14	5.15	5.15	5.15
14	4.21	4.42	4.55	4.63	4.70	4.78	4.83	4.87	4.91	4.96	5.00	5.04	5.06	5.07	5.07	5.07
15	4.17	4.37	4.50	4.58	4.64	4.72	4.77	4.81	4.84	4.90	4.94	4.97	4.99	5.00	5.00	5.00
16	4.13	4.34	4.45	4.54	4.60	4.67	4.72	4.76	4.79	4.84	4.88	4.91	4.93	4.94	4.94	4.94
17	4.10	4.30	4.41	4.50	4.56	4.63	4.68	4.72	4.75	4.80	4.83	4.86	4.88	4.89	4.89	4.89
18	4.07	4.27	4.38	4.46	4.53	4.59	4.64	4.68	4.71	4.76	4.79	4.82	4.84	4.85	4.85	4.85
19	4.05	4.24	4.35	4.43	4.50	4.56	4.61	4.64	4.67	4.72	4.76	4.79	4.81	4.82	4.82	4.82
20	4.02	4.22	4.33	4.40	4.47	4.53	4.58	4.61	4.65	4.69	4.73	4.76	4.78	4.79	4.79	4.79
22	3.99	4.17	4.28	4.36	4.42	4.48	4.53	4.57	4.60	4.65	4.68	4.71	4.74	4.75	4.75	4.75
24	3.96	4.14	4.24	4.33	4.39	4.44	4.49	4.53	4.57	4.62	4.64	4.67	4.70	4.72	4.74	4.74
26	3.93	4.11	4.21	4.30	4.36	4.41	4.46	4.50	4.53	4.58	4.62	4.65	4.67	4.69	4.73	4.73
28	3.91	4.08	4.18	4.28	4.34	4.39	4.43	4.47	4.51	4.56	4.60	4.62	4.65	4.67	4.72	4.72
30	3.89	4.06	4.16	4.22	4.32	4.36	4.41	4.45	4.48	4.54	4.58	4.61	4.63	4.65	4.71	4.71
40	3.82	3.99	4.10	4.17	4.24	4.30	4.34	4.37	4.41	4.46	4.51	4.54	4.57	4.59	4.69	4.69
60	3.76	3.92	4.03	4.12	4.17	4.23	4.27	4.31	4.34	4.39	4.44	4.47	4.50	4.53	4.66	4.66
100	3.71	3.86	3.98	4.06	4.11	4.17	4.21	4.25	4.29	4.35	4.38	4.42	4.45	4.48	4.64	4.65
∞	3.64	3.80	3.90	3.98	4.04	4.09	4.14	4.17	4.20	4.26	4.31	4.34	4.38	4.41	4.60	4.68

*Using special protection levels based on degrees of freedom.

Source; Duncan, D. B., Biometrices, Vol II, 1955

Multiple range table value [at the 5% level for 16 (n_2) and 4 or 3 or 2(p)] multiplied by the standard error is shown in the following table:

	Mean or Least Square Means			
	10.8	10.4	9.0	5.4
5.4	3.23X.58=1.87	3.15X.58=1.83	3.00X.58=1.74	-
9.0	3.15X.58=1.83	3.00X.58=1.74	-	
10.4	3.00X.58=1.74	-		
10.8	-			

These values are then compared to the difference between mean values as follows:

	Mean or Least Square Means			
	10.8	10.4	9.0	5.4
5.4	5.4 (1.87)	5.0 (1.83)	3.6 (1.74)	-
9.0	1.8 (1.83)	1.4 (1.74)	-	
10.4	0.4 (1.74)	-		
10.8	-			

Comparing the difference between means with the table values shows the following difference value to be larger than the table value and therefore significantly different at the 5% level (* symbol used).

	Mean or Least Square Means			
	10.8	10.4	9.0	5.4
5.4	*	*	*	-
9.0	NS	NS	-	
10.4	NS	-		
10.8	-			

Significance can be shown by lining means up in order of increasing or decreasing magnitude and drawing a line under the ones that are not significantly different as follows:

<u>10.8 10.4 9.0</u> 5.4

This would indicate that the 5.4 mean (group 4) is significantly different from all the other groups (Group 1, 2, 3). The other groups (1, 2, 3) are not significantly different from each other and for example treatments 2 and 3 difference in tenderization may be due to chance alone. However the tenderizer in group 4 did make the product significantly more tender.

Analysis of variance for a regression problem:

In a regression problem the significance of the regression line was previously checked by using sample size to determine the required correlation values for significance. This significance can also be evaluated by using analysis of variance. In the "cold cut - thickness vs. weight problem" found in Chapter IX, the following calculations were made:

1	2	3	4	5	7	8	9	10	12
Sample	Thickness in millimeters X	Average thickness in millimeters $\bar{X} = \frac{\sum X}{n}$	Deviation from the average thickness $x = (X-\bar{X})$	Deviation from the average thickness squared $x^2 = (X-\bar{X})^2$	Weight in grams Y	Average weight in grams $\bar{Y} = \frac{\sum Y}{n}$	Deviation from the average weight $y = (Y-\bar{Y})$	Deviation from the average weight squared $y2 = (Y-\bar{Y})2$	Cross Products xy or $(X-\bar{X})\cdot(Y-\bar{Y})$
A	4	4.0	0	+ 0	33	37.0	- 4.0	+ 16.0	- 0.0
B	2	4.0	-2.0	+4.0	16	37.0	-21.0	+441.0	+42.0
C	3	4.0	-1.0	+1.0	30	37.0	- 7.0	+ 49.0	+ 7.0
D	6	4.0	+2.0	+4.0	62	37.0	+25.0	+625.0	+50.0
E	5	4.0	+1.0	+1.0	44	37.0	+ 7.0	+ 49.0	+ 7.0
n=5	$\sum X=20$		$\sum x=0$	$\sum x^2=10.0$	$\sum Y=185$		$\sum y=0$	$\sum y^2=1180.0$	$\sum xy=106.0$

Additional calculations required are as follows:

Deviation from Regression Calculation

X Thickness in millimeters	Y Weight in grams	$\hat{Y} = -5.4 + 10.6X$ Regression equation for predicting Y values from X values	$d = Y-\hat{Y}$ Difference between observed Y values & predicted Y value	$(Y-\hat{Y})^2$ d^2 Differences squared	$y^2 = (Y-\bar{Y})^2$ Deviation from average wt. squared See Chapter IX
4	33	37.0	-4.0	16.00	16.0
2	16	15.8	0.2	0.04	441.0
3	30	26.4	3.6	12.96	49.0
6	62	58.2	3.8	14.44	625.0
5	44	47.6	-3.6	12.96	49.0
			$\Sigma = 0$	$\Sigma d^2 = 56.40$ deviation from regression	$\Sigma y^2 = 1180.0$

Total variation calculation (Total Sum of Squares)

$\Sigma y^2 = \Sigma(Y-\bar{Y})^2 = 1180$ from previous table (or Chapter IX)

r^2 = Proportion of variation in Y accounted for by X

$1-r^2$ = Proportion of variation in Y not accounted for by X

Σy^2 = Sum of deviation from average $\bar{Y}$ squared (total variation)

Σd^2 = Sum of deviation from predicted $\hat{Y}$ squared

Reduction in variation due to regression calculation

$$r^2\,\Sigma y^2 = (.9758)^2\ (1180) = (.9522)\ (1180) = 1123.6$$

r^2 = Proportion of variation in Y accounted for by X

Σy^2 = Sum of deviation from average $\bar{Y}$ squared (total variation)

or

$\hat{Y}$	$\bar{Y}$	$\hat{Y}-\bar{Y}$	$(\hat{Y}-\bar{Y})^2$
37	37	0	0
15.8	37	-21.2	449.44
26.4	37	-10.6	112.36
58.2	37	21.2	449.44
47.6	37	10.6	112.36
		Σ	1123.60

Check on reduction in variation due to regression calculation

$$\text{Check on } r^2\,\Sigma y^2 = \Sigma(\hat{Y}-\bar{Y})^2 = \Sigma y^2 - \Sigma 2^d = 1180 - 56.4 = 1123.6$$

$\Sigma(\hat{Y}-\bar{Y})^2$ = Variability accounted for by regression lin

Σy^2 = Total variation

Σd^2 = Variability not accounted for by regression line

<u>Check on Σd^2</u> [Variability not accounted for by regression line (sum of squares of deviation from regression)]

$$\Sigma d^2 = (1-r^2)(\Sigma y^2)$$
$$= [1-(.9758)^2] \bullet [1180]$$
$$= [1-.9522 \quad] \bullet [1180]$$
$$= (.0478) \bullet (1180) = 56.40$$ (see previous table)

<u>or</u>

$$\Sigma d^2 = \Sigma(Y-\hat{Y})^2 = \text{see table} = 56.40$$

<u>Degrees of freedom calculation</u>

total = n - 1 = 5 - 1 = 4

regression = 1

deviation from regression = n - 2 = 5 - 2 = 3 or

total df - regression = 4 - 1 = 3

<u>Sum of Squares Explanation for Regression</u>

The same line of reasoning can be used to explain the "Sum of Squares" for the regression problem that was used in the preceding analysis of variance problem.

The "total sum of squares" is obtained by subtracting the Y-value for each observation from the overall Y mean value ($\bar{Y}$) and this deviation ($Y-\bar{Y}$) is squared and summed [$\Sigma(Y-\bar{Y})^2$]. This represents [$\Sigma(Y-\bar{Y})^2$] the total Y deviation in this set of data. See calculation in previous table.

The "total sum of squares" representing the total deviation in the data is then segmented into the deviation of "Regression" (can be assigned to regression line) and the "deviation from regression" (not explained by regression line).

To calculate the "regression" sum of squares the predicted (regression line) Y value ($\hat{Y}$) is substracted from the average Y-value ($\bar{Y}$) for each observation and this deviation ($\hat{Y}-\bar{Y}$) is squared and summed [$\Sigma(\hat{Y}-\bar{Y})^2$]. This [$\Sigma(\hat{Y}-\bar{Y})^2$] represents the portion of the total deviation that is accounted for by the regression line. See calculation in previous table.

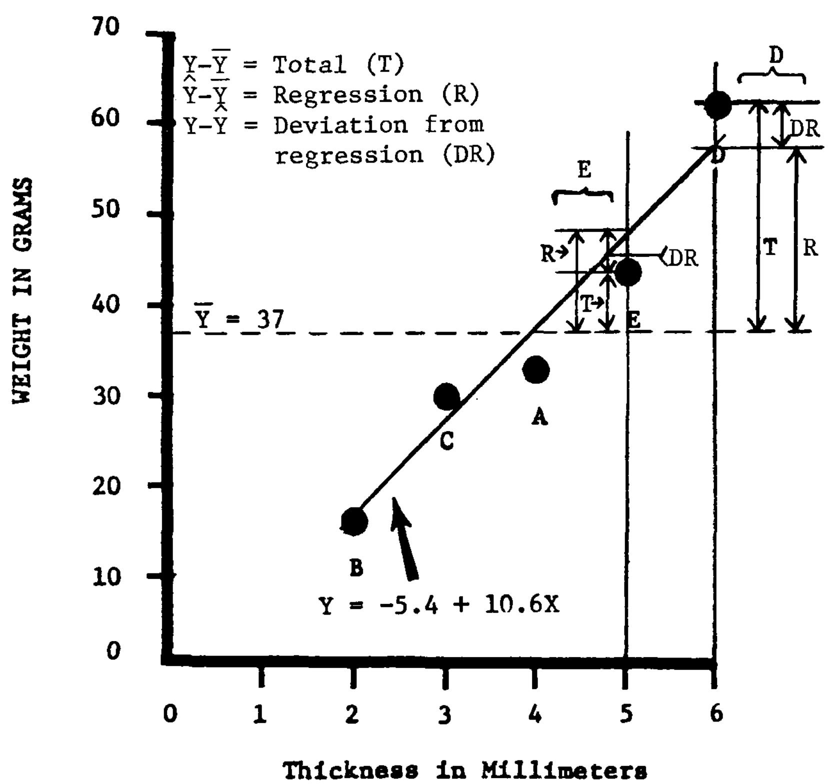
The "deviation from regression" sum of squares is the portion of the "total sume of squares" not accounted for by regression. It can be obtained by subtracting the "regression sum of squares" $[\Sigma(\hat{Y}-\bar{Y})^2]$ from the "total sum of squares" $[\Sigma(Y-\bar{Y})^2]$ or can be calculated by subtracting the Y value for each observation from the predicted (regression line) Y value $(\hat{Y})$ for each observation $(Y-\hat{Y})$. This deviation $(Y-\hat{Y})$ is then squares and summed $[\Sigma(Y-\hat{Y})^2]$. This represents the portion of the "total sum of squares" not accounted for by the regression line. See calculation in preceding table.

To summarize these individual deviations they are grouped in the following table for each observation. Notice that the deviation due to regression plus the deviation from regression equals the total deviation.

Sample	X	Y	$\hat{Y}$	$\bar{Y}$	Total deviation $(Y-\bar{Y})$	Deviation due to regression $(\hat{Y}-\bar{Y})$	Deviation from regression $(Y-\hat{Y})$	$(\hat{Y}-\bar{Y}) + (Y-\hat{Y})$
A	4	33	37	37	-4	0	-4	-4
B	2	16	15.8	37	-21	-21.2	0.2	-21
C	3	30	26.4	37	-7	-10.6	3.6	-7
D	6	62	58.2	37	+25	21.2	3.8	25
E	5	44	47.6	37	+7	10.6	-3.6	7
		ΣY=185			Σ=0	Σ=0	Σ=0	

$$\bar{Y}=\frac{\Sigma Y}{n} = \frac{185}{5} = 37$$

To diagramatically illustrate these deviations they are drawn vertically on the following graph for samples D and E.



WEIGHT IN GRAMS
70
60
50
40
30
20
10
0
Y-Ȳ = Total (T)
Ŷ-Ȳ = Regression (R)
Y-Ŷ = Deviation from regression (DR)
Ȳ = 37
D
DR
E
R
DR
T
E
T
R
D
C
A
B
Y = -5.4 + 10.6X
0
1
2
3
4
5
6
Thickness in Millimeters

Analysis of variance table

Sources of Variation	Sum of squares	Degrees of freedom (df)	Mean squares or variance estimate	"F" ratio
Regression	$r^2 \Sigma y^2 = \Sigma(\hat{Y}-\overline{Y})^2 = 1123.6$	1	$\left[\frac{\text{Sum of Squares}}{\text{df}}\right]$ 1123.6	$\left[\frac{\text{Mean square regression}}{\text{Mean square deviation}}\right]$ $\frac{1123.6}{18.8} = 59.77$
Deviation from regression	$(1-r^2) \Sigma y^2$ or Σd^2 or $\Sigma (Y-\hat{Y})^2 = 56.40$	n-2 = 3	$\left[\frac{\text{Sum of Squares}}{\text{df}}\right]$ 18.8	
Total	$\Sigma y^2 = \Sigma(Y-\overline{Y})^2 = 1180$	n-1 = 4		

The calculated "F" ratio is 59.77 with 1-3 degrees of freedom. The table "F" values using the above degrees of freedom (1-3) and the 95% probability level is 10.13 and 34.12 using the 99% probability level.

This would indicate that the regression line is significant at the 99% level (agrees with correlation - significance interpretation) unless the one chance out of 100 has occurred in which an "F" value is expected to exceed 34.12 (Not very probable).

$$r^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} \times 100$$

$$= \frac{1123.6}{1180} \times 100 = 95.22\%$$

= 95.22% is the percent of the total sum of squares that is accounted for by the regression equation.

This agrees with the correlation squared value

$r^2 = (.9758)^2$

= 95.22% of Y accounted for by X

Sample Problems

1. Calculate the following by using the t-test.

X
1
2
4

A. Calculate the sample mean and check the results.
B. Calculate the sample standard error.
C. Using the t-test determine if this sample could have been drawn for a population with a mean of 4. The standard t value for 3 degrees of freedom at the 95% probability level is 3.18.
D. Calculate the 95% confidence interval for the mean of the above sample.

2. Calculate the following by using the analysis of variance.

X_1	X_2	X_3
1	1	3
3	3	3
2	5	6
6	9	12

A. Calculate the mean of X_1, X_2, and X_3 and check the calculation.

B. Calculate n_1, n_2, n_3 and n total.

C. Calculate $\sum X$ total.

D. Calculate $\sum X^2$ total.

E. Calculate $\sum x^2$ total.

F. Calculate $\sum x^2$ between groups.

G. Calculate $\sum x^2$ within group.

H. Check $\sum x^2$ within group.

I. Fill in an "Analysis of Variance" table.

J. Calculate the F ratio.

K. The table F Value at 95% level for 2, 6 degrees of freedom is 5.14.

Were X_1, X_2, and X_3 significantly different?

Answers (A-2,3,4, B-3,3,9, C-27, D-103, E-22, F-7, G-15, J-1.4, K-No.)

3. Calculate the following by using the analysis of variance and the Duncan Multiple Range Test.

X_1	X_2	X_3
2	2	4
2	2	6
1	2	5

A. Calculate the mean of X_1, X_2 and X_3 and check the calculation.

B. Calculate n_1, n_2, n_3 and n total.

C. Calculate ΣX_{total}

D. Calculate ΣX^2_{total}

E. Calculate Σx^2_{total}

F. Calculate $\Sigma x^2_{between\ groups}$

G. Calculate $\Sigma x^2_{within\ group}$

H. Check $\Sigma x^2_{within\ group}$

I. Fill in analysis of variance table

J. Calculate the F ratio

K. What F values are required to be significant at the 95 & 99% level?

L. Is X_1, X_2, X_3 significantly different? At what level?

M. By using the Duncan Multiple Range Test which means are different?

Answers (A-1.67, 2.00, 5.00, B-3, 3, 3, 9, C-26, D-98, E-22.889, F-20.222, G-2.667, H-2.667, I-table, J-22.747, K-5.14, 10.92, L-yes, 99% level, M-5.00, 2.00, 1.67.

References

Haber, Audrey and Richard P. Runyon. 1969. General Statistics, Addison-Wesley Publishing Company, Reading, Massachusetts, Menlo Park, California, London, Don Mills, Ontario.

Kramer, C. Y. 1956. Extension of Multiple Range test to groups means with unequal number of replications. Biometrics 12. 307.

Snedecor, George W. 1956. Statistical Methods. The Iowa State College Press, Ames, Iowa.

Steel, Robert G. D. and James H. Torrie. 1960. Principles and Procedures of Statistics. McGraw-Hill Book Company, Inc., New York, Toronto, London.

Accepting or Rejecting

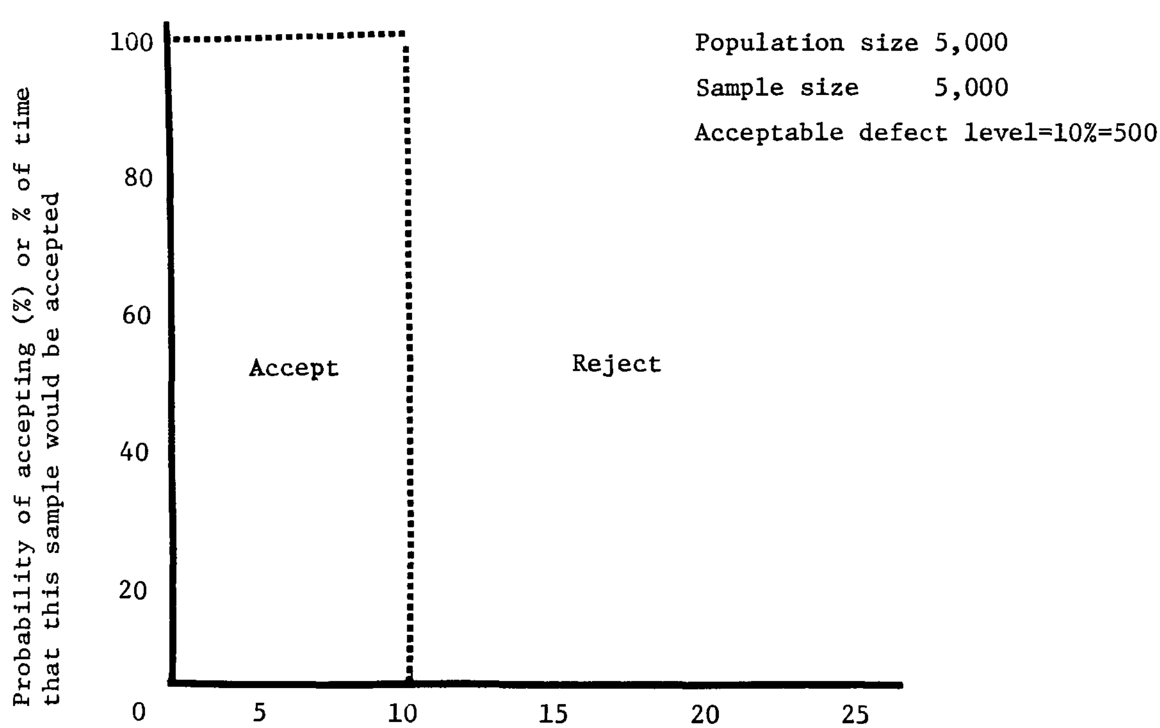
Populations need to be accepted or rejected when material changes or is about to change hands. For example, when raw material that have been purchased arrives at the dock or when a manufactured product is completed and is ready for shipment or when a product or process is evaluated by a federal inspector.

Decision on whether to accept or reject a population are usually based on results obtained from a sample.

If a sample contains all (the total) of the observations in a population and an excellent job of evaluation is performed on the observations then the sampling plan will always make the correct decision about the population.

If we would plot an operating curve (OC) of this situation it would appear as follows:

Sampling plan that would always correctly (ideal operating curve) evaluate the population. Producers and consumers risk would be as low as possible.



% Defective in the sample or average number of defects/100 units. [In this case (sample size=population size) it is also the % defectives in the population].

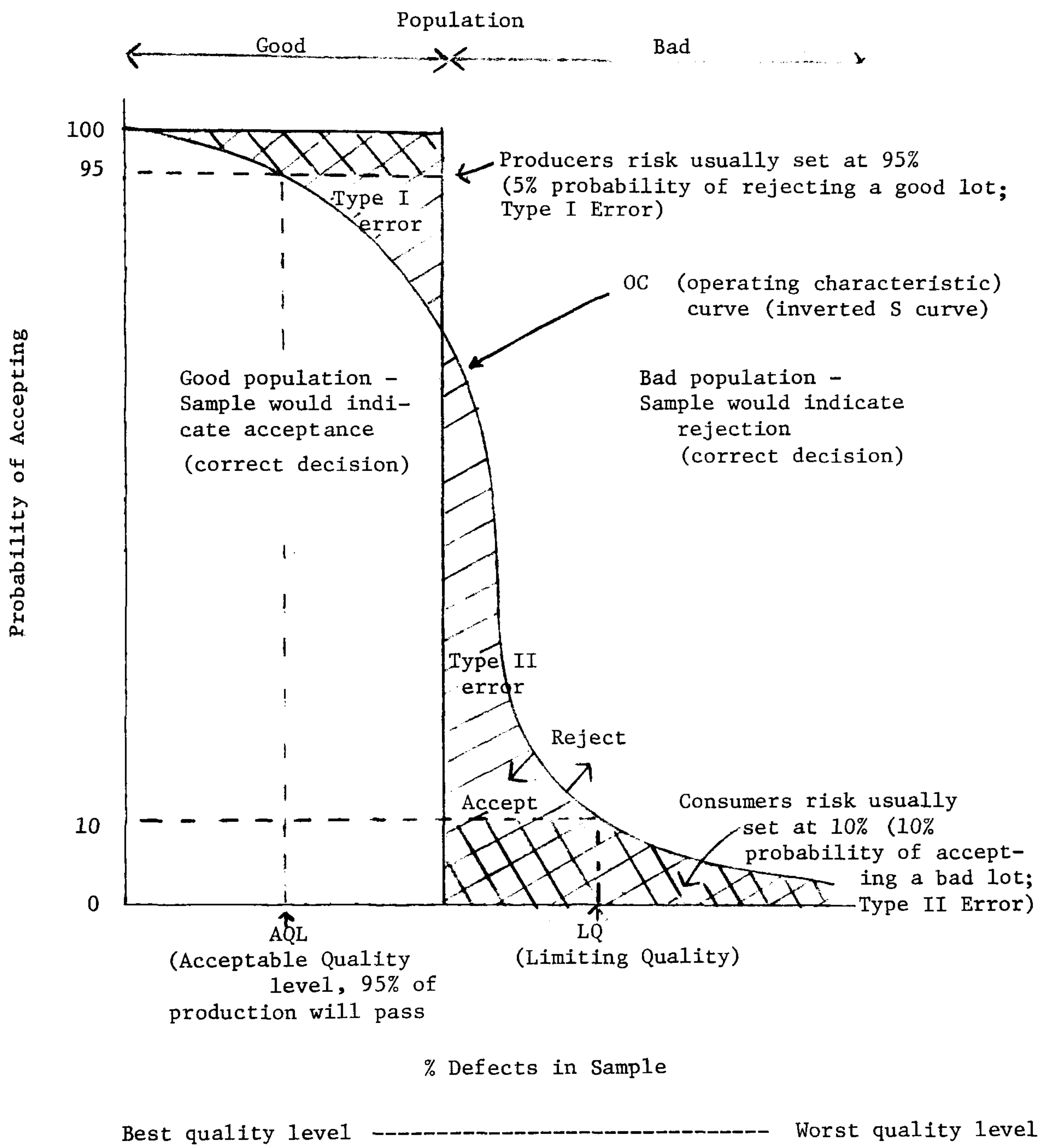
The operating characteristic curve is the communicating link between the sampling plan numbers and the lot to be inspected since it shows the discriminatory powers of a sampling plan.

Usually only a portion (incomplete information) of the population is observed as a sample and the actual quality of the population is unknown. Under these conditions a decision must be made based on the sample and there are 4 possible combinations of results. These are:

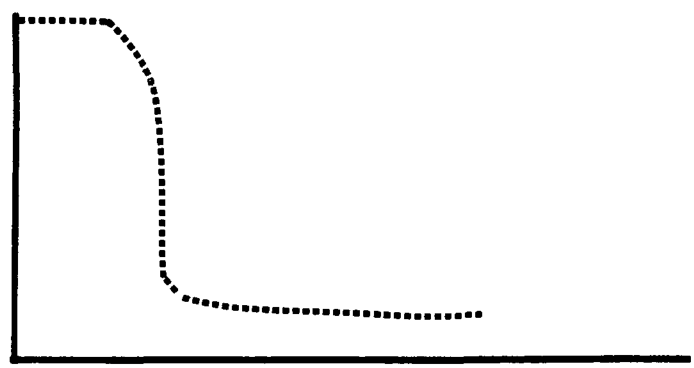
	True situation of the population	
	Good	Bad
Decision arrived at by examining the sample		
Accept	Correct decision	Type II error
Reject	Type I error	Correct decision

Every sampling plan has an O C curve.

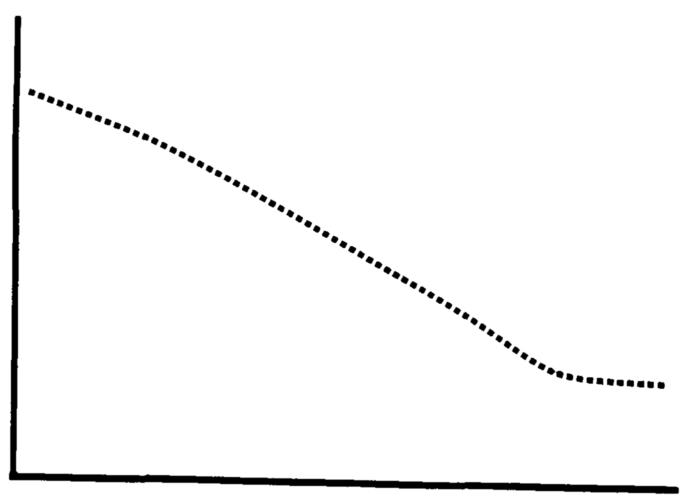
When incomplete information is available the sample operating curve (OC) deviates from ideal operating curve and the sample operating curve superimposed on the ideal operating curve would look as follows:



Therefore the steeper the sample operating curve (or the more nearly it approaches the ideal operating curve) the greater will be its accuracy and less errors will be committed.



The flatter the operating curve the greater the number of errors and the accuracy will be diminished.

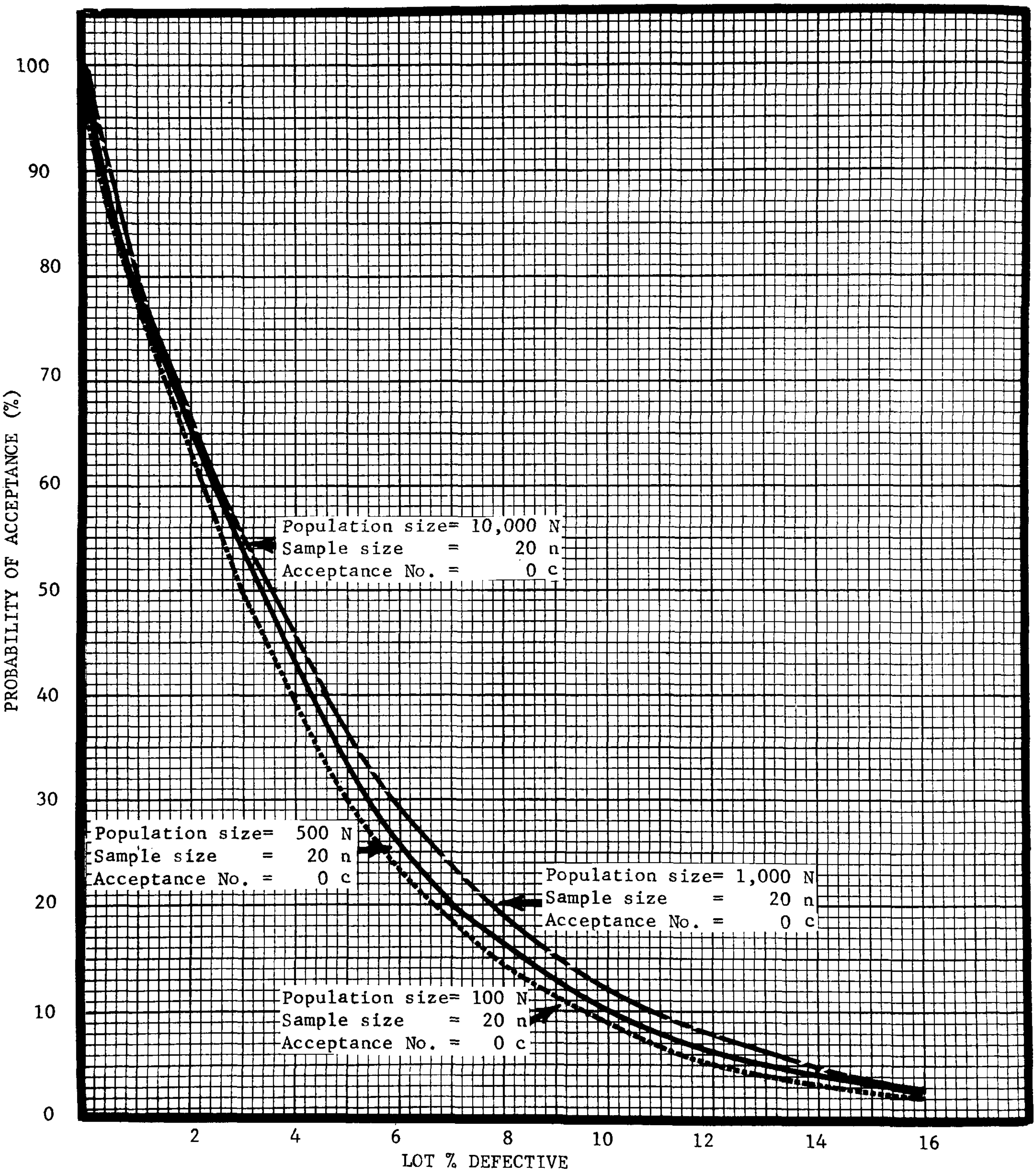


Three things (independently or in combination) will influence the shape of an operating curve.

1. Population size (N)
2. Sample size (n)
3. Acceptance number (c) or maximum number of defects.

An example of how these influence the operating curves are shown as follows:

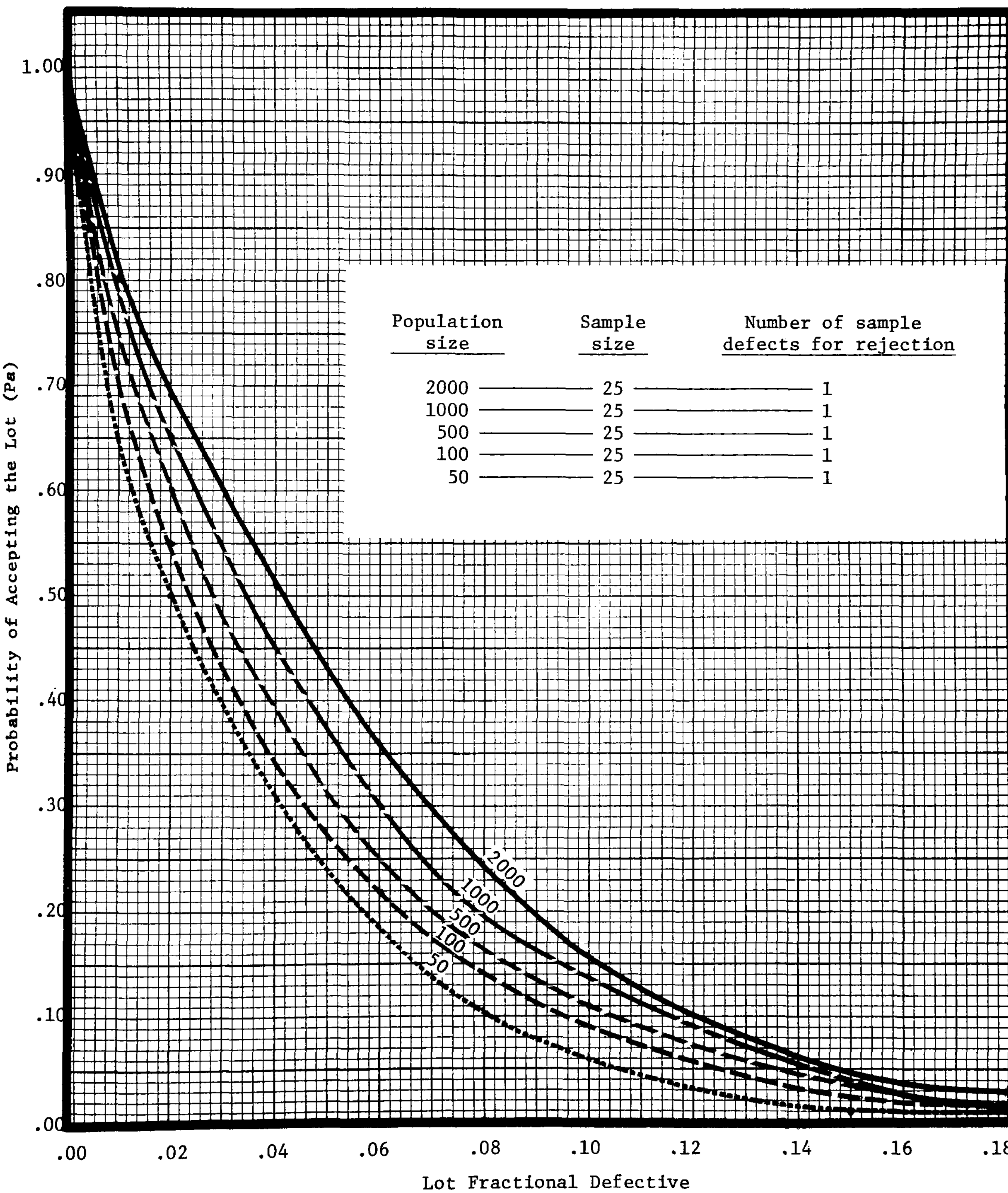
Effect of Changing the Population Size and Maintaining a Constant Sample Size and Acceptance Number



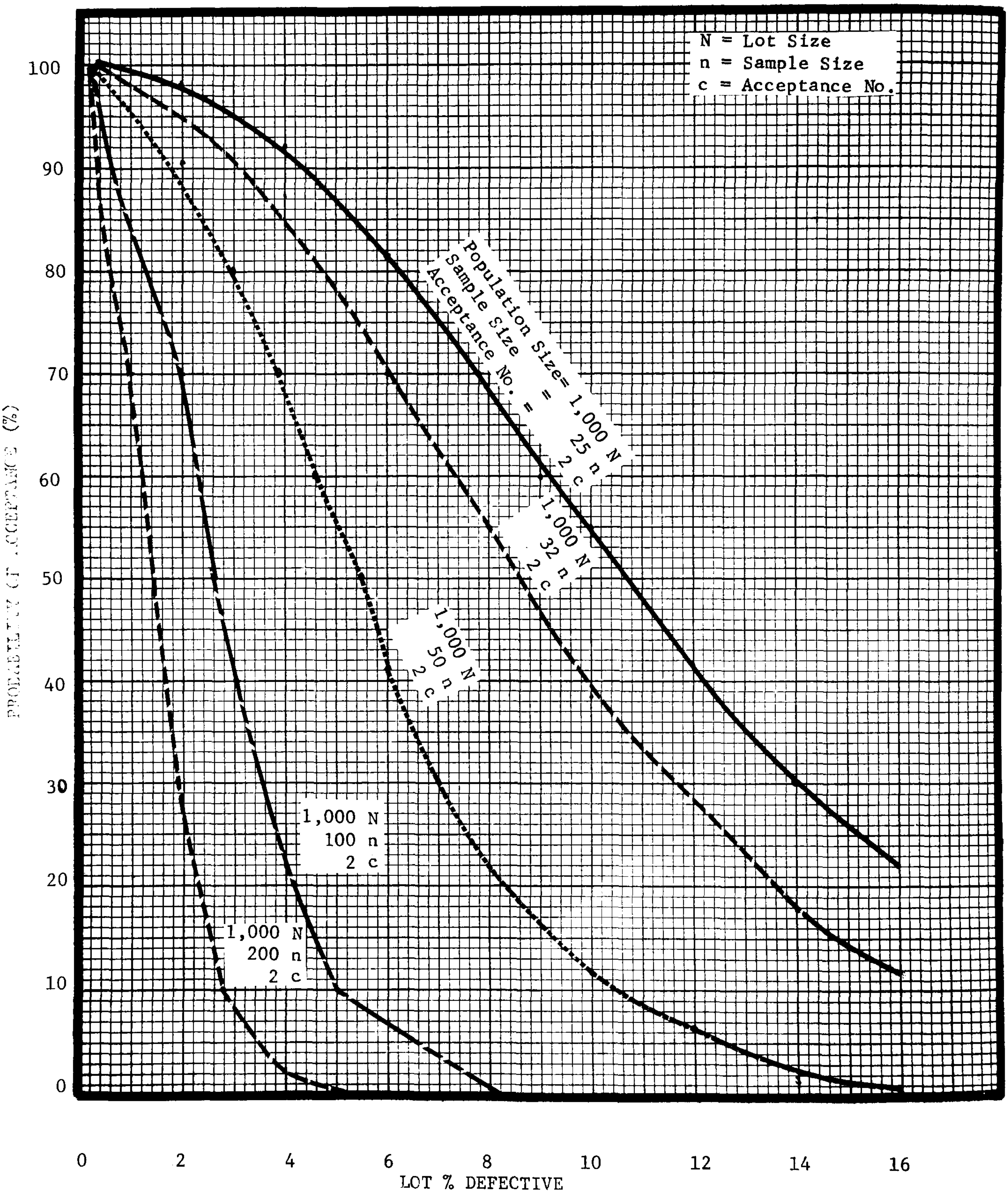
A CONSTANT SAMPLE SIZE GIVES A FAIRLY CONSTANT AMOUNT OF PROTECTION REGARDLESS OF THE POPULATION SIZE

An example of constant sample size type of operating curve constructed from repeated sample of meat products is illustrated as follows:

Operating Characteristic Curves Single Sample of 25, Acceptance Number, c =0
Variable Population (2000, 1000, 500, 100, 50)

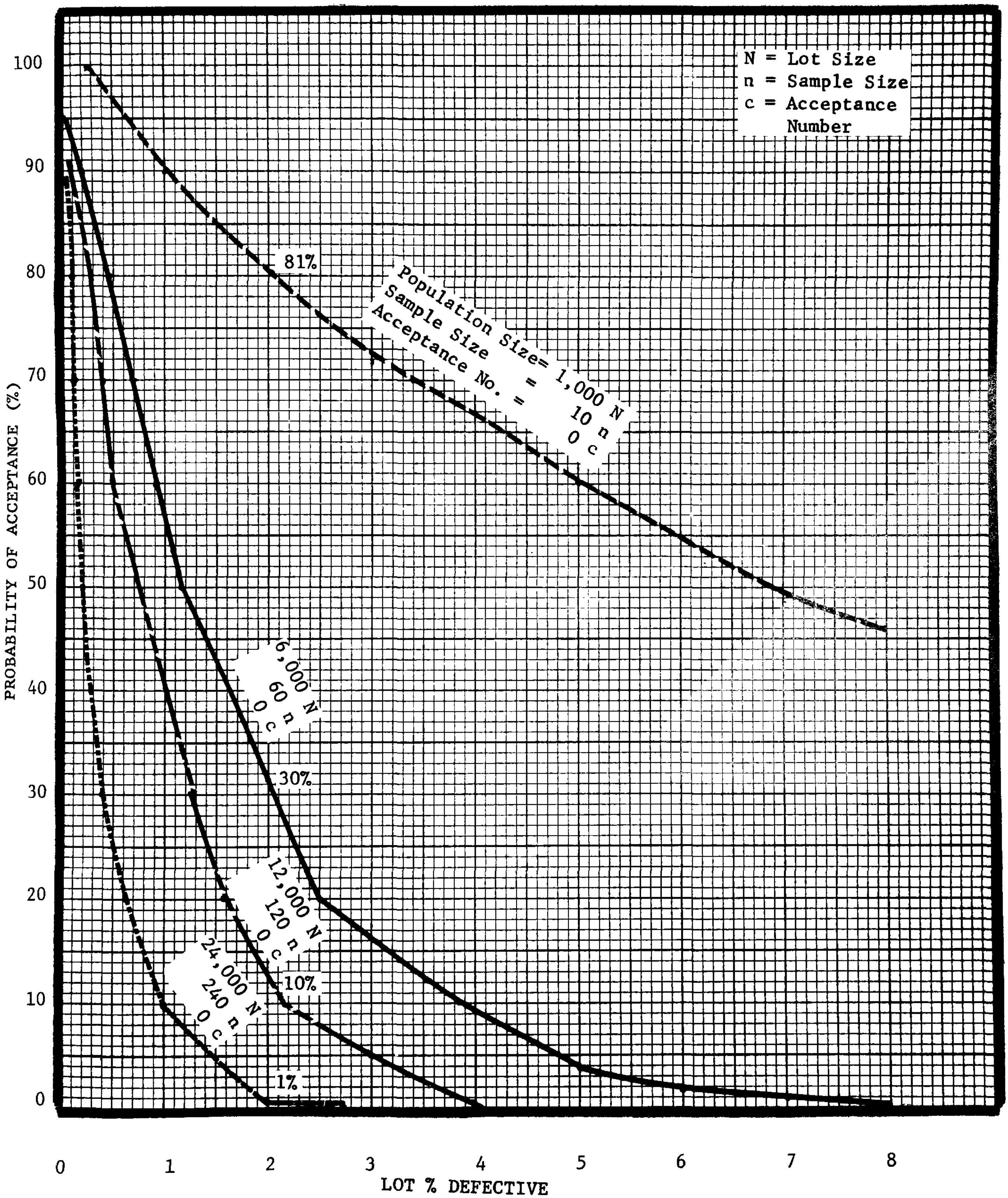


Effect of Changing the Sample Size and Maintaining a Constant Population Size and Acceptance Number



INCREASING THE SAMPLE SIZE WHILE THE POPULATION SIZE REMAINS CONSTANT GIVES INCREASED PROTECTION

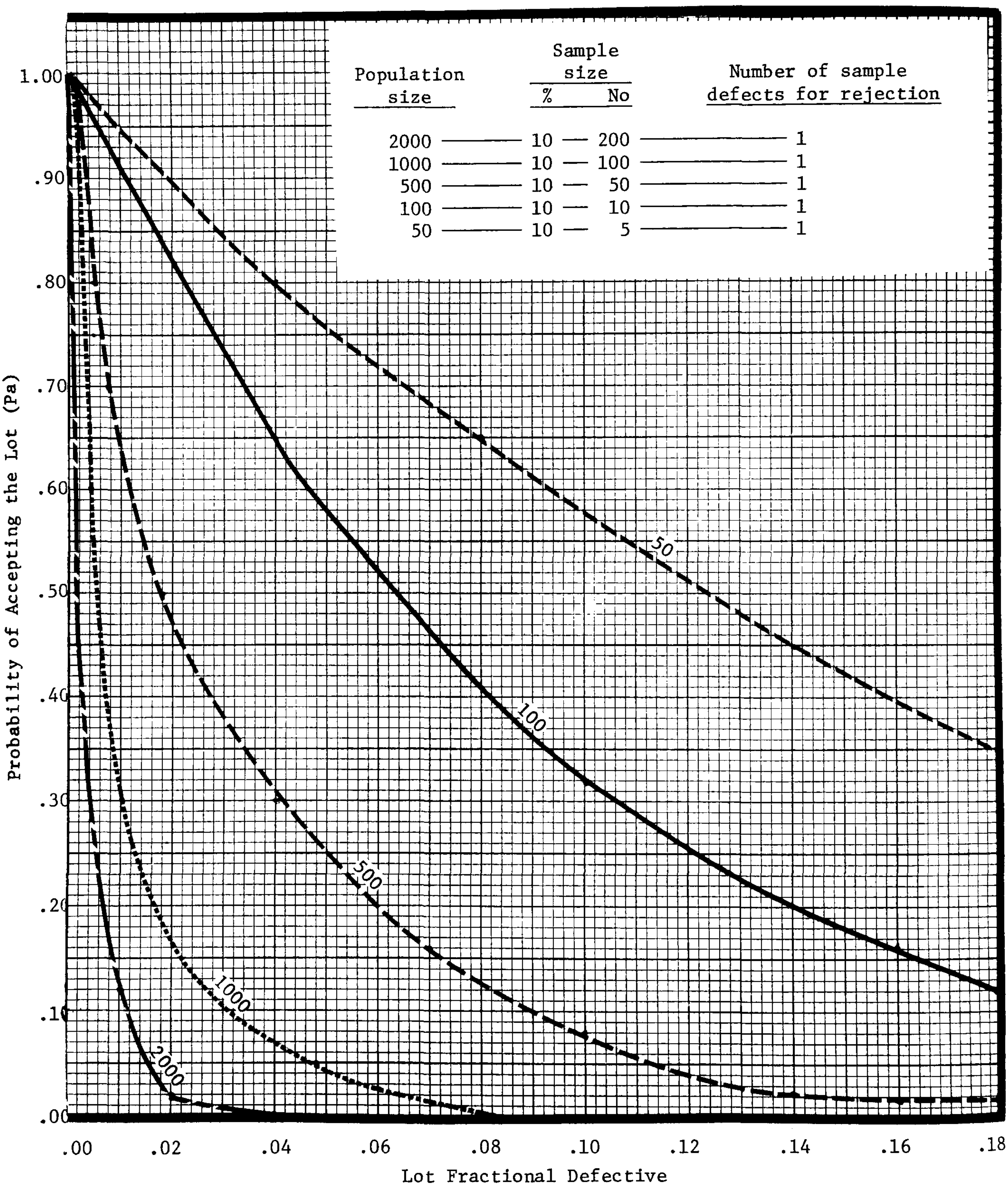
Percent Sampling Has Both a Variable Population Size and a Variable Sample Size But a Constant Acceptance Number



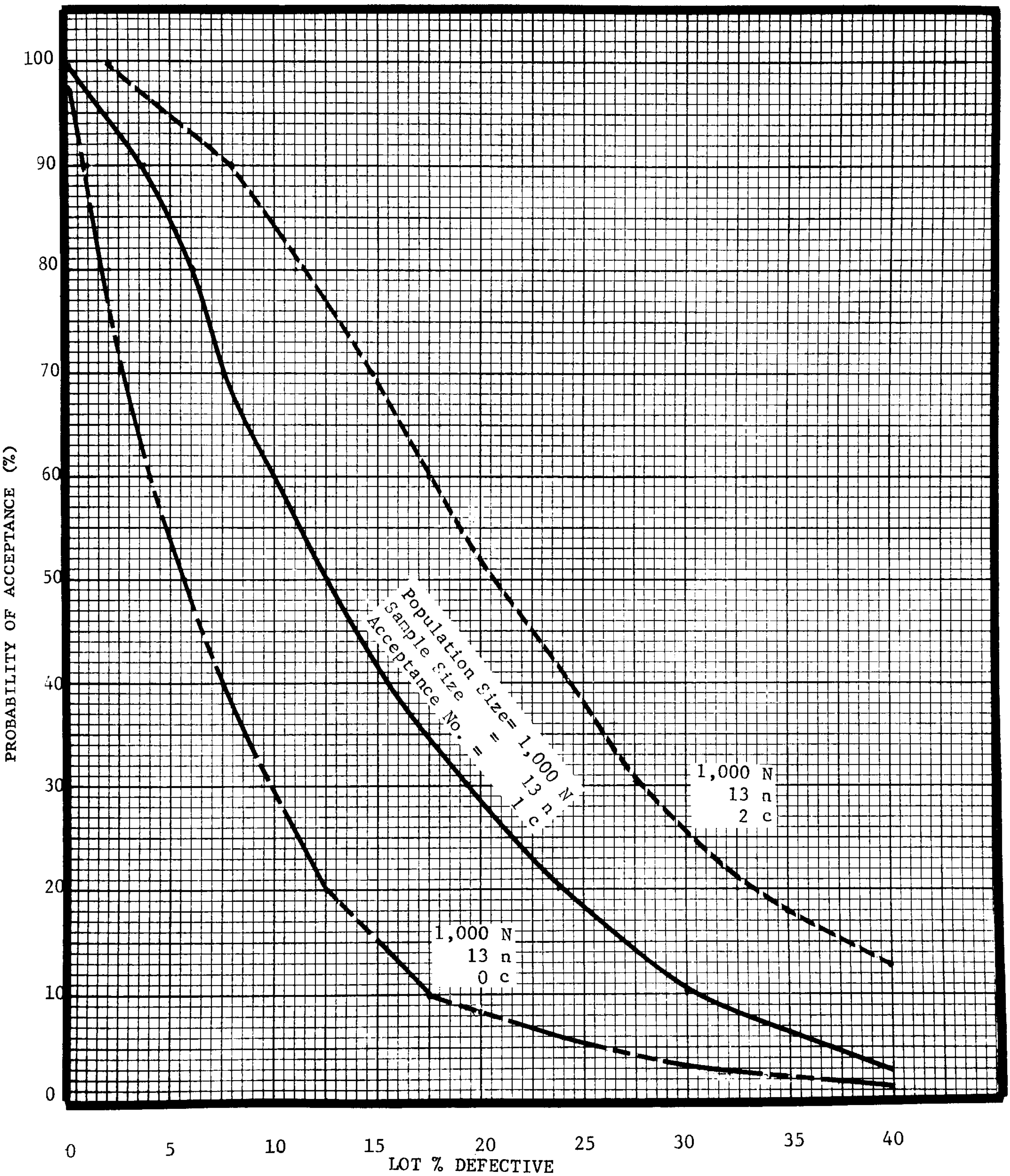
THE SAME % OF SAMPLES GIVES DIFFERENT PROTECTION. THEREFORE, PERCENT SAMPLING IS NOT A DESIRABLE PROCEDURE

An example of percentage sampling type of operating curve constructed from repeated samples of meat products is illustrated as follows:

Operating Characteristic Curves Sample Size of 10%, Acceptance Number, c=0
Variable Population (2000, 1000, 500, 100, 50)

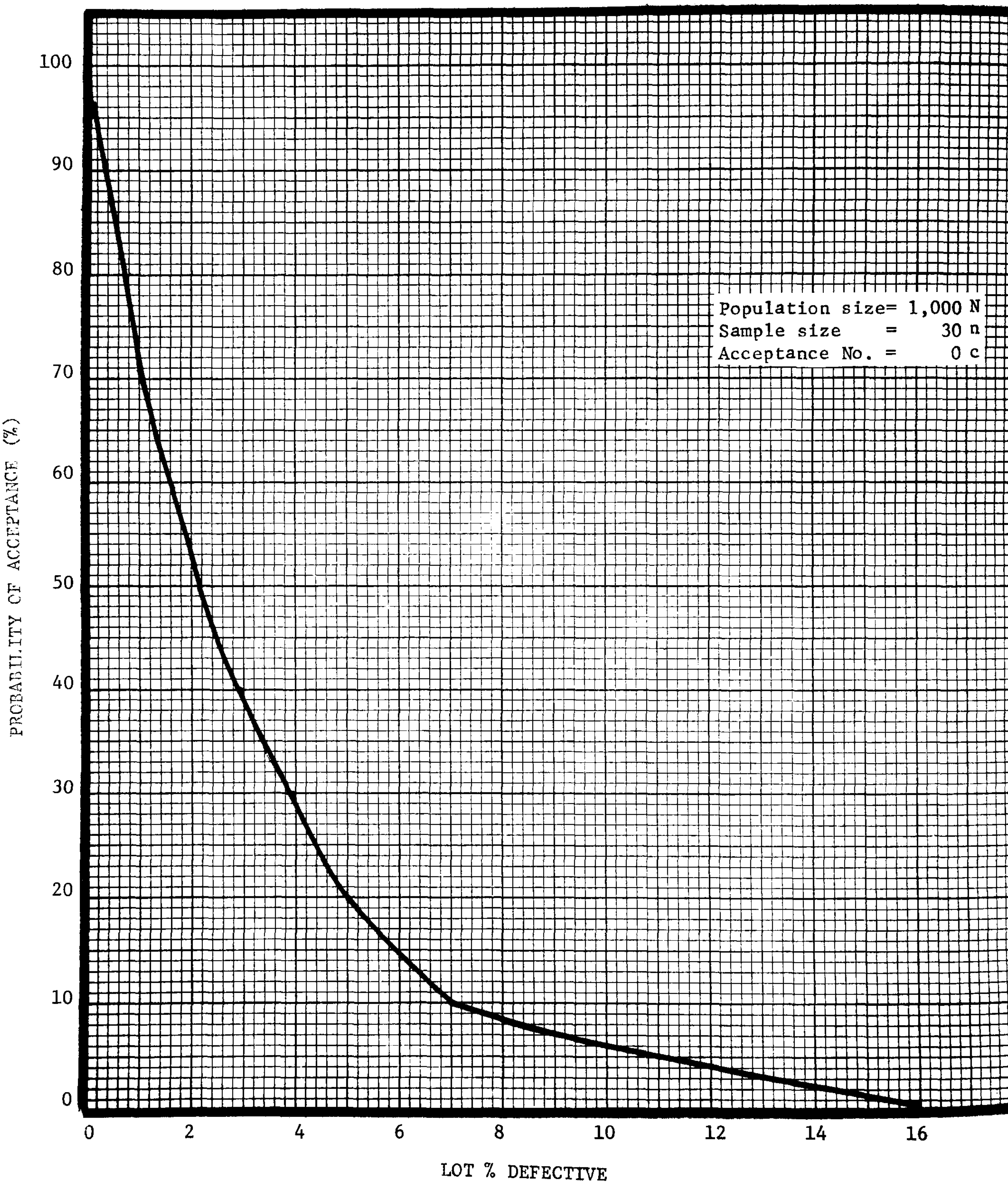


EFFECT OF CHANGING THE ACCEPTANCE NUMBER & MAINTAINING A CONSTANT POPULATION AND SAMPLE SIZE

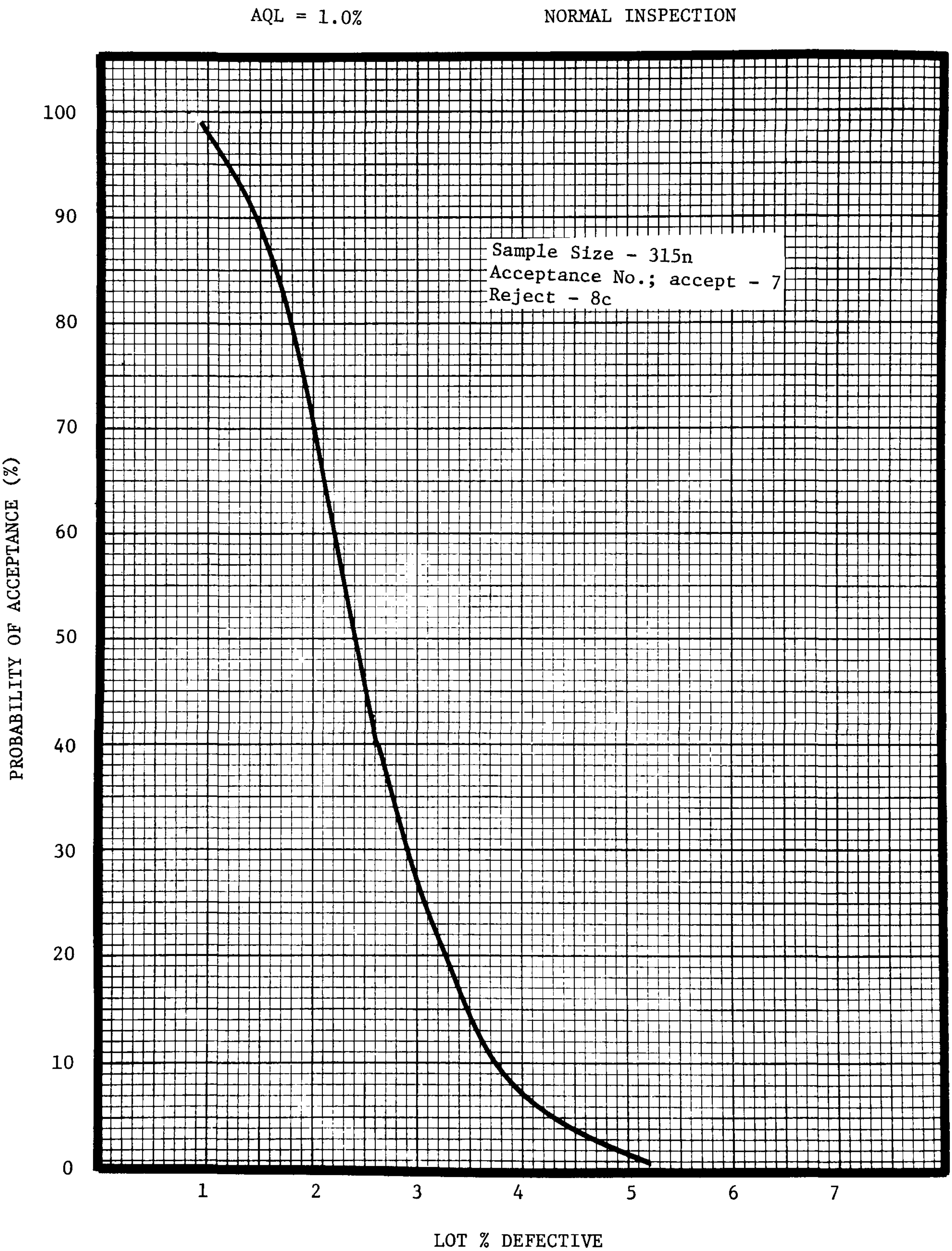


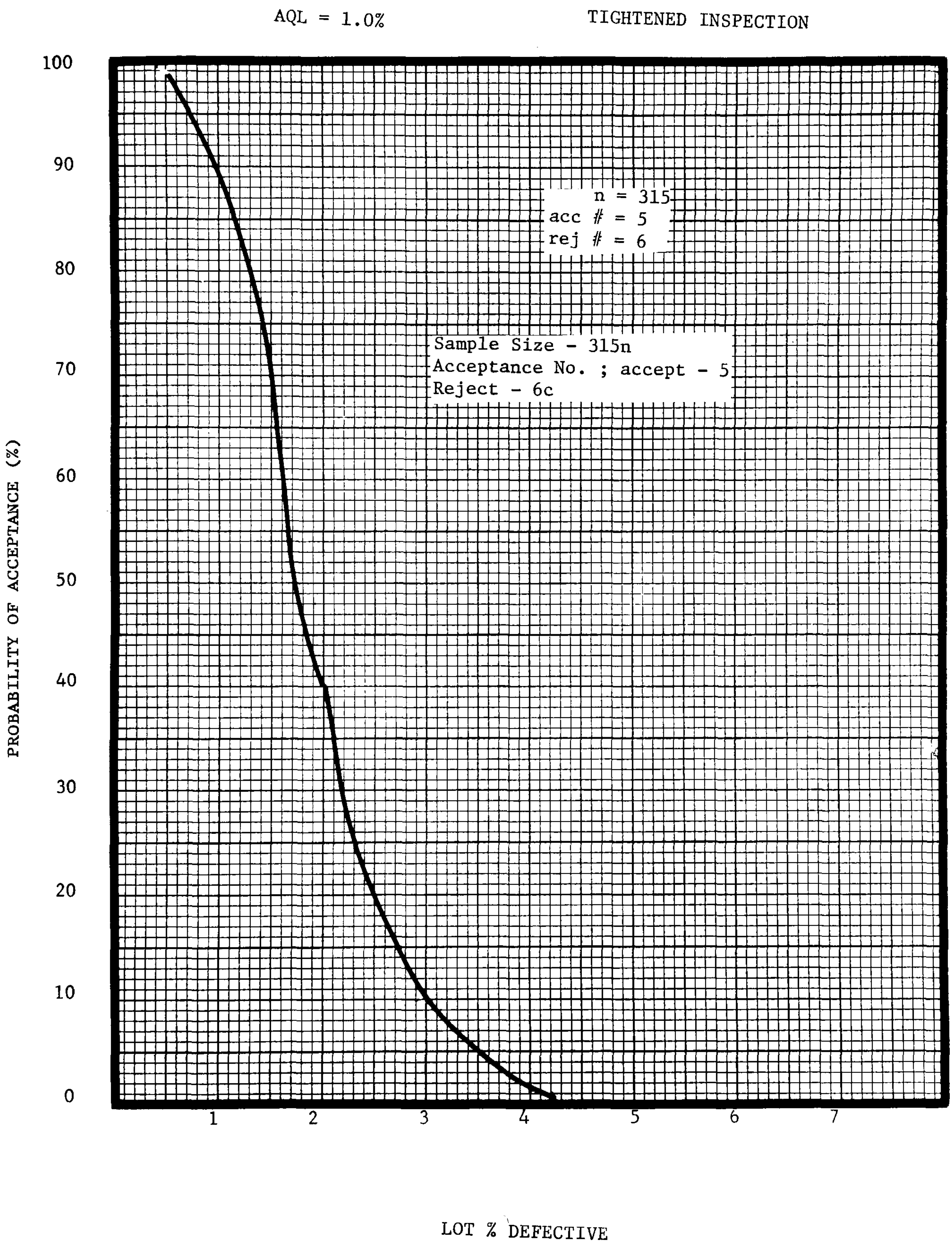
INCREASING THE ACCEPTANCE NUMBER MOVES THE OC CURVE TO THE RIGHT WHICH INCREASES THE PROBABILITY OF ACCEPTANCE AT ALL LOT PERCENT DEFECTIVE LEVELS

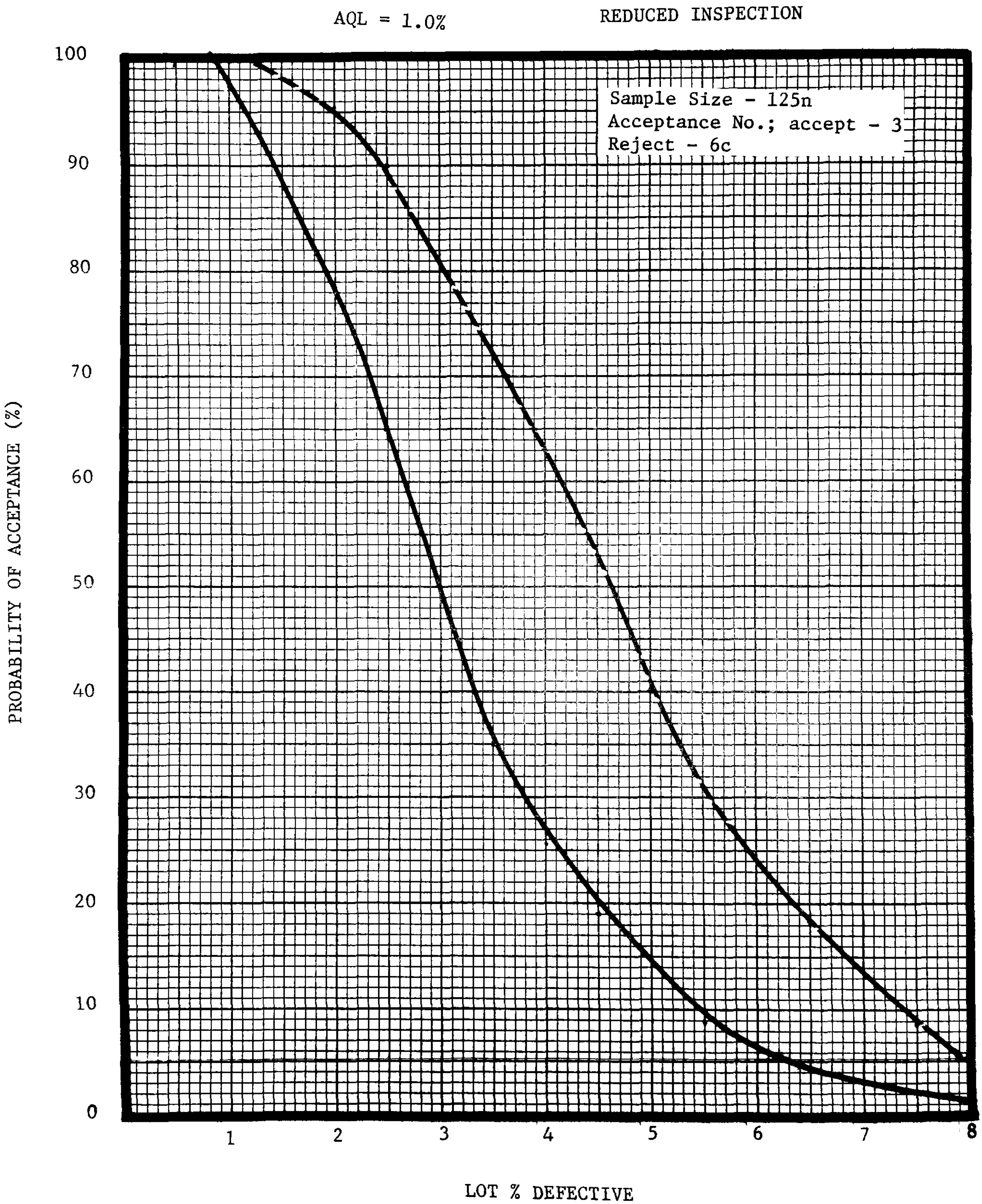
AN ACCEPTANCE LEVEL OF 0 DOES NOT ASSURE THAT THE ACCEPTED POPULATION WILL BE PERFECT



Changing inspection seventy also changes the operating curve as shown on the next three graphs of normal inspection, tightened inspection and reduced inspection.


AQL = 1.0%
NORMAL INSPECTION
Sample Size - 315n
Acceptance No.; accept - 7
Reject - 8c
PROBABILITY OF ACCEPTANCE (%)
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
LOT % DEFECTIVE


AQL = 1.0%
TIGHTENED INSPECTION
n = 315
acc # = 5
rej # = 6
Sample Size - 315n
Acceptance No. ; accept - 5
Reject - 6c
100
90
80
70
60
50
40
30
20
10
0
PROBABILITY OF ACCEPTANCE (%)
1
2
3
4
5
6
7
LOT % DEFECTIVE


AQL = 1.0%
REDUCED INSPECTION
Sample Size - 125n
Acceptance No.; accept - 3
Reject - 6c
PROBABILITY OF ACCEPTANCE (%)
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
LOT % DEFECTIVE

SAMPLING

The foundation on which most statistical problems are based is the selection of a representative sample from the population. No conclusions are valid unless the sampling portion of the experiment is properly conducted. The now famous computer terminology of GIGO which means:

"Garbage In"

"Garbage Out"

can very readily be applied to proper sampling. If sampling is improperly conducted this is certainly "garbage in" to the statistical analysis portion of the problem.

Errors in Sampling

Two of the most common errors in sampling may be outlined as follows:

1. Convenient sample - There is always the temptation to select the sample that is the most convenient. The box on the bottom of the stack certainly is more difficult to sample than the box on the top of the stack. Product produced in the middle of the day is more convenient to sample than product produced at an odd hour.

   Since the result of all statistical analyses is to predict parameters of a population, then all samples of the population should have an equal likelihood of being selected for the sample--even the box on the bottom of the stack or the product produced at a late hour.

2. Judgment sample - Often an individual doing the sampling thinks that by his knowledge of the product he can select a representative sample from the population. Individuals close to the population may have information that will be quite valuable in subdividing these

populations (will be discussed later) but selecting samples by this method is not a good practice. It is a rare individual that has no opinion (even subconsciously) on the outcome of a statistical problem in which he is sampling and, therefore, there is a subconscious tendency to select a sample that will favor this outcome. If this takes place then every sample in the population did not have the same likelihood of being chosen.

Random Sampling

This method of sampling states that all observations in the population have an equal likelihood of being selected for the sample. This selection can be accomplished by several techniques as follows:

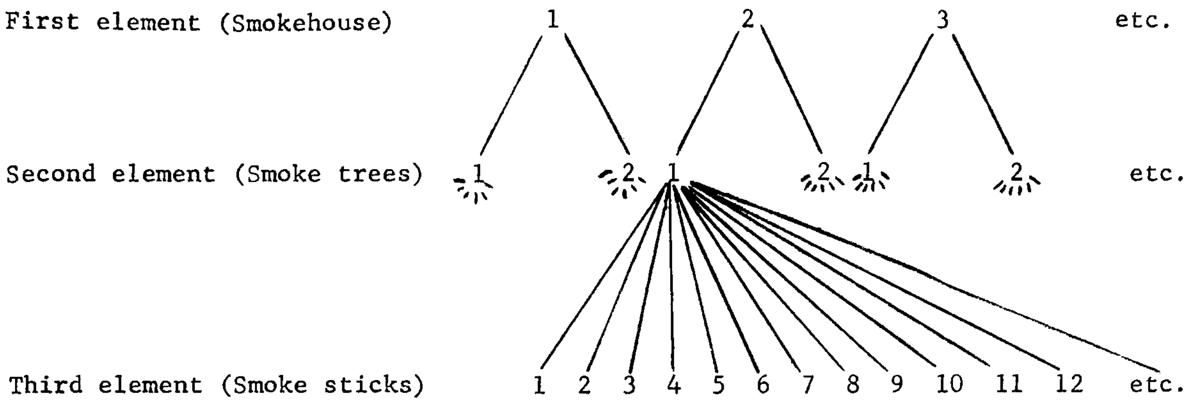
1. List all possible samples on a piece of paper, shuffle the samples or mix them in a hat and randomly draw the desired number of samples. This procedure for large populations requires a considerable quantity of effort and, therefore, it is very seldom used.

2. Random table (the 100% inspection table in Chapter 1 can be used) - There are random digit tables that have already shuffled the digits and randomly selected and listed them. The table may be entered at any point and by proceeding in any direction the digits taken will have been randomly selected.

   This means that if all units in the population are assigned a number and a random digit table is used to pick the samples, then the individual observations each have the same likelihood of being chosen.

Element Sampling

If each sample can be classified by the elements of its construction then the sampling procedure may be simplified. For example, the following is the daily wiener production of a plant:



For this sampling procedure the smokehouses to be used are randomly selected. The smoke trees are then randomly selected from each of the smokehouses chosen. The smoke sticks are randomly selected from each of the smoke trees chosen. In this type of sampling each smoke stick had an equal chance of being selected but the sampling problem was much simpler than if all the original smoke sticks had been enumerated.

<u>Previous Knowledge of Problem Used in Sampling</u>

1. <u>Stratified Sampling</u>: Often some information is available on the population to be sampled and this information can be used to improve the sampling procedure. This procedure may be outlined as follows:

   a. The population is subdivided into groups which are mutually exclusive and exhaustive. The subdividing should be done so that the different groups have different means and unless this is done there is no advantage in this procedure. Also the fraction of the total population that falls into each subgroup should be known.

      An example of how this sampling procedure might be used is shown as follows:

   An examination is being made on the cooler shrink of beef carcasses in Plant A. This plant has two chill coolers and due to their construction these coolers have different chilling capacities, relative humidity and air flow values. In this case the mean shrink values would be expected to be different in the individual coolers and a stratified sampling plan should be followed. Each cooler would be designated as a subgroup of the total production of Plant A.

b. Random samples would be selected from each subgroup for examination.

In the above examples carcasses would be selected from each cooler.

c. Estimates are made for each subgroup.

In the above example a mean shrink value would be calculated for each cooler. This information is of only secondary interest in predicting the total population shrink value but later may become of primary interest if cooler modifications are to be made.

d. The subgroup estimates are combined to obtain a total population estimate and are normally weighted according to the fraction of the total population contained in each subgroup.

In the above example the average shrink in each cooler would be weighted by the fraction of the plant carcasses that are chilled in this cooler.

2. Severity of Inspection: Previous information concerning a product is not always used to physically change the sampling plan, however, this information is often used to change the required performance of an acceptable sample. This can be accomplished in a statistics problem by changing the acceptable probability level (Example, changing table "t" or "F" values from a probability level of 95% to 99%). In actual practice almost the same thing is accomplished by changing the requirements a sample must meet to be considered acceptable. Levels of acceptability are often divided as follows:

a. Normal inspection. This inspection is usually used at the start of production when no information about the product is available and is continued when there is no evidence that the product is appreciably better or poorer than specified.

b. Tightened inspection. This inspection is used when there is evidence that the product is deteriorating. Acceptance criteria are more stringent than under normal inspection and these acceptance criteria are continued until there is evidence that the product has improved.

c. Reduced inspection (often not used and never without appropriate authority). This inspection may be used when the product consistently meets or exceeds specifications. If the product deteriorates then normal inspection is again used.

Systematic Sampling

This sampling plan is used when there is some natural arrangement of samples. This arrangement should not be related to the value being measured and should be random in nature.

An example would be the weight of packages of cold cuts coming from a packaging line.

The first sample to be chosen is randomly selected and then every n'th sample is chosen.

This plan has the advantage of evenly spreading the samples through the total population, but care must be taken to see that there is no cyclic variation in production that corresponds to the n'th sampling selection or the conclusions drawn will not be valid.

This type of plan is often used to monitor a production line and when the values start to drift then corrective actions are taken.

Sampling Plans

There are three basic sampling plans that differ only in the number of samples needed in order to reach a decision. These are divided as follows:

1. Single: A lot is inspected and the results either meet the specifications and the lot is accepted or the lot fails to meet the specifications and the lot is rejected.
2. Double: In this sampling plan there is a range between where a lot is accepted and where it is rejected. If the results of the first sample are not good enough to accept the lot and not bad enough to reject the lot then a second sample is drawn. The results of the total of the first and second samples are used to accept or reject the lot. The advantage of this plan is that good and bad lots are detected by the first sample and that only lots near the specification level require both the first and second samples to be made.
3. Multiple: Continuation of double sampling plan and more than two samples are required to reach a decision.

Size of Sample

In statistical quality control or inspection sampling, the size of the sample is often specified. How was this sample size determined or how would you obtain the sample size if you were initiating a statistical problem? The question of sample size is often difficult to answer but very important for the following reasons:

1. Sample too small: A sample that is too small may fail to detect a difference that is important. The accuracy of an estimate of a population parameter is related to the size of the sample with small samples yielding less accurate estimates.
2. Sample too large: Samples cost time and money. Therefore, too large a sample may be a waste of both. Since the accuracy of an estimate of a population parameter is related to the size of the sample, too large a sample may give a more accurate estimate than is required.

In order to determine the sample size to use, information is needed in the following two areas.

1. <u>An estimate of the standard deviation of the population</u> (σ): The less variability within a population (measured by σ) the more accurate a sample will estimate the population parameter. The more variability, the less accurate this estimate will be. Since increasing sample size will also increase accuracy of this estimate it follows that a population with increased variability will need a larger size sample to achieve the same degree of accuracy of estimation as a population with less variability.

   The estimate of a standard deviation of the population (σ) may be obtained from previous experiments on this type of population or from some knowledge of the population. For example, the approximate range of a population may be known (or estimated) and from the normal curve (Chapter VI) it was shown that 6 standard deviations would include 99.73% of the total observations. Therefore, the following relationships would yield an estimate of the standard deviation of the population.

$$\text{Crude estimate of } \sigma = \frac{\text{range of a large number of observations from population}}{6}$$

2. <u>Smallest difference between the sample statistic and the population parameter that will be detected</u>: The smaller the sample size becomes, the larger the error in estimating the population parameter. The question then becomes how big an error will you tolerate? This is going to depend on the following:

   a. What use is to be made of the estimate? Will you only adjust future production (more error acceptable) or will you retain & rework present production (less error acceptable).

b. What is the consequence of committing an error? Will an error cause an economic problem (more error acceptable) or a health problem (less error acceptable).

In the final analysis the acceptable error is an arbitrary decision based on the above information remembering that additional security is paid for by a larger sample size. When the allowable error is determined it should be expressed as shown in the following terms:

L = Allowable error between sample statistic mean and population parameter mean.

Percent of time (determined from the number of standard deviations) that the error will exceed the L value.

Formula for calculation of sample size

$$n = \frac{4\sigma^2}{L^2}$$

n = sample size
4 = 2^2 = 2 standard deviations or 95% probability level or 5% chance that the error will exceed L (use 3^2 or 9 for 99% level)
σ = Estimate of standard deviation of the population.
L = Allowable error between sample statistic mean and population parameter mean.

Sample Problem

Sample for percent fat in a frankfurter product made in the plant described in Chapter VIII under the subheading "Description of a Population".

Information available

sample, s = 0.6 which is an estimate of the standard deviation of the population (σ).

sample range = 30.9 - 28.5 = 2.4

The population range would probably exceed this value since this is only a small sample (21 observations). By checking other analysis (500 observations) of this same product a high value of 31.2 and a low value of 28.1 is found.

range of population = 31.2 - 28.1 = 3.1

crude estimate of $\sigma = \frac{3.1}{6} = 0.5$

This gives 2 estimates of the standard deviations of this population of 0.6 and 0.5 (the larger estimate of 0.6 will be used because it is probably more accurate and it will necessitate taking a larger sample).

Decisions to be made

How close must the sample mean percent fat agree with the population mean percent fat? With analysis of raw materials the percent fat can only be controlled within 0.5% in this plant; therefore, 0.5% is used as an acceptable error.

Ninety-five percent confidence interval will be used so that the error will exceed 0.5% fat only 5% of the time.

Calculation of Sample Size

$$n = \frac{4\sigma^2}{L^2}$$

$$n = \frac{4X(.6)^2}{(.5)^2} = \frac{4X.36}{.25} = \frac{1.44}{.25} = 5.76 \text{ or 6 samples}$$

The mean of 6 samples drawn from this population should agree within 0.5% of the fat content of the population mean 95% of the time.

"Rule of thumb" based on the above equations can be stated as follows:

95% probability level with an allowable error of 0.5

$n = (\sigma \text{ X } 4)^2$

99% probability level with an allowable error of 0.5

$n = (\sigma \text{ X} 6)^2$

Physically Selecting the Samples

Regardless of the sampling plan chosen there will be judgment decisions to be made by the individual selecting the sample. This individual should keep in mind that the samples should be random and the samples are to be used as predictors of the total population.

References

ASTM. 1958. Choice of Sample Size to Estimate the Average Quality of a Lot or Process. ASTM Designation E122 - 58.

Cochran, William G. and Gertrude M. Cox. 1957. Experimental Designs. John Wiley & Sons, Inc., New York, London, Sydney.

Department of the Army. 1969. Food Inspection Specialist. TM8 - 451 Headquarters, Dept. of the Army, Washington, D. C.

McCarthy, Philip J. 1957. Introduction to Statistical Reasoning. McGraw-Hill Book Company, Inc., New York, Toronto, London.

Spiegel, Murray R. 1961. Theory and Problems of Statistics. McGraw-Hill Book Company, New York, St. Louis, San Francisco, Toronto, Sydney.

Snedecor, G. W. 1956. Statistical Methods. The Iowa State College Press, Ames, Iowa.

Walker, Helen M. and Joseph Lev. 1958. Elementary Statistical Methods. Holt, Rinehart and Winston, New York.

SYMBOLS & FORMULAS

$=$	Equal
$\neq$	Not equal
$x < y$	x is less than y
$x > y$	x is greater than y
$\leq$	less than or equal to
$\geq$	greater than or equal to
$\sqrt{}$	square root
$\sum$	sum of all values that follow
$\sum_{i=1}^{n} Xi$	sum of all quantitites from X_1 through X_n
	Example $X_1 + X_2 + X_3 \ldots . X_n$
μ	population mean
σ^2	population variance
σ	population standard deviation
b	slope of line in a regression equation $b = \dfrac{n\sum XY - (\sum X)(\sum X)}{n\sum X^2 - (\sum X)^2}$
a	constant in regression equation $a = \overline{Y} - b\overline{X}$
f	Frequency
n	total number of quantities
df	degrees of freedom n-1
r	correlation coefficient $r = \dfrac{\sum xy}{\sqrt{\sum x^2 \ \sum y^2}} = \dfrac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$
R	range = Large X - Small x
$\overline{R}$	mean range $\overline{R} = \dfrac{\sum R}{n}$
MR	moving range $X_i - X_i + 1$
$\overline{MR}$	mean moving range $\overline{MR} = \dfrac{\sum MR}{n-1}$

$\bar{p}$	p bar	$\bar{p} = \frac{\sum h}{\sum n}$
s^2	sample variance	
s	sample standard deviation	$s = \sqrt{\sum x^2/n}$ $\quad s = \sqrt{\frac{\sum X^2}{n} - \bar{X}^2}$
X	(any capital letter) variable	
$\bar{X}$	Arithmetic means	$\bar{X} = \frac{\sum X}{n} = \frac{\sum fX}{n}$
$\bar{X}$-s	Bar X, standard deviation quality control charts	
$\bar{X}$-r	Bar X, Range Quality Control Charts	
L	difference between sample statistic mean and population parameter mean	
$\sum$ x	(any small leter) deviation of a variable from the mean for that variable;	

$$x = X - \bar{X}$$

$$\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

$\sum X^2$	sum of individual scores after they have been squared
$(\sum X)^2$	values are summed first and the total squared
∞	Infinity

FINDING THE SQUARE ROOT

Method:

1. Find the decimal point and mark off digits in groups of two in both directions starting at the decimal point. Add zeros to the left of the decimal point to achieve an even number of digits. Also add zeros to the right of the decimal to achieve twice the number of digits that are desired to the right of the decimal in the answer.

2. Mark the decimal point for the answer above the decimal point in the number whose square root is to be taken. There will be a digit in the answer for each pair of digits in the original number.

3. Find the largest perfect square contained in the extreme left hand group.

4. Place the square root in the answer and the square below the 1st digit or pair of digits. Subtract the square from the 1st digit or pair of digits.

5. Bring down the next pair of digits.

6. Double the partial answer.

7. Add a trial digit to the right of the "double partial answer". Multiply this new number by the trial digit. Place the correct new number in the answer.

8. Subtract the product. (Occasionally this number will be larger than the "doubled partial answer plus the new digit" but the next highest digit will make the product too large).

9. Repeat steps 5, 6, 7, 8 until the desired accuracy is achieved.

10. Check the answer by:

    a.) If the square is perfect -

    Multiply the square root by itself and the answer should be the original number.

    b.) If the square is not perfect -

    Find the square root to the desired number of decimal places. Round the last decimal place as desired. Multiply the square root by itself and if the square root was rounded down then add the remainder. (If square root was rounded up then subtract the overage) and the answer should be the original number.

Example: Square root of 32 to 3 decimal places.

```
            5. 6 5 7                      Check -    5.657
          √32.000000                                 5.657
            25                                       39599
        106 | 700                                   28285
              636                                  33942
      1125 |  6400                                28285
               5625                             32.001649
    11307  |   77500                               - 1649
                79149                           32.000000
                -1649
```

Find the square root of the following numbers:

A. 940900.00

B. 781

C. 620

D. 270400

E. .4

Check your answers.

Answers (A-970, B-27.946, C-24.900, D-520., E-.6325)

Table of Squares and Square Roots of Numbers

From 1 to 1000

Number	Square	Square root	Number	Square	Square root
1	1	1.000	51	2601	7.141
2	4	1.414	52	2704	7.211
3	9	1.732	53	2809	7.280
4	16	2.000	54	2916	7.348
5	25	2.236	55	3025	7.416
6	36	2.449	56	3136	7.483
7	49	2.646	57	3249	7.550
8	64	2.828	58	3364	7.616
9	81	3.000	59	3481	7.681
10	100	3.162	60	3600	7.746
11	121	3.317	61	3721	7.810
12	144	3.464	62	3844	7.874
13	169	3.606	63	3969	7.937
14	196	3.742	64	4096	8.000
15	225	3.873	65	4225	8.062
16	256	4.000	66	4356	8.124
17	289	4.123	67	4489	8.185
18	324	4.243	68	4624	8.246
19	361	4.359	69	4761	8.307
20	400	4.472	70	4900	8.367
21	441	4.583	71	5041	8.426
22	484	4.690	72	5184	8.485
23	529	4.796	73	5329	8.544
24	576	4.899	74	5476	8.602
25	625	5.000	75	5625	8.660
26	676	5.099	76	5776	8.718
27	729	5.196	77	5929	8.775
28	784	5.292	78	6084	8.832
29	841	5.385	79	6241	8.888
30	900	5.477	80	6400	8.944
31	961	5.568	81	6561	9.000
32	1024	5.657	82	6724	9.055
33	1089	5.745	83	6889	9.110
34	1156	5.831	84	7056	9.165
35	1225	5.916	85	7225	9.220
36	1296	6.000	86	7396	9.274
37	1369	6.083	87	7569	9.327
38	1444	6.164	88	7744	9.381
39	1521	6.245	89	7921	9.434
40	1600	6.325	90	8100	9.487
41	1681	6.403	91	8281	9.539
42	1764	6.481	92	8464	9.592
43	1849	6.557	93	8649	9.644
44	1936	6.633	94	8836	9.695
45	2025	6.708	95	9025	9.747
46	2116	6.782	96	9216	9.798
47	2209	6.856	97	9409	9.849
48	2304	6.928	98	9604	9.899
49	2401	7.000	99	9801	9.950
50	2500	7.071	100	10000	10.000

Table of Squares and Square Roots (Continued)

Number	Square	Square root	Number	Square	Square root
101	10201	10.050	151	22801	12.288
102	10404	10.100	152	23104	12.329
103	10609	10.149	153	23409	12.369
104	10816	10.198	154	23716	12.410
105	11025	10.247	155	24025	12.450
106	11236	10.296	156	24336	12.490
107	11449	10.344	157	24649	12.530
108	11664	10.392	158	24964	12.570
109	11881	10.440	159	25281	12.610
110	12100	10.488	160	25600	12.649
111	12321	10.536	161	25921	12.689
112	12544	10.583	162	26244	12.728
113	12769	10.630	163	26569	12.767
114	12996	10.677	164	26896	12.806
115	13225	10.724	165	27225	12.845
116	13456	10.770	166	27556	12.884
117	13689	10.817	167	27889	12.923
118	13924	10.863	168	28224	12.961
119	14161	10.909	169	28561	13.000
120	14400	10.954	170	28900	13.038
121	14641	11.000	171	29241	13.077
122	14884	11.045	172	29584	13.115
123	15129	11.091	173	29929	13.153
124	15376	11.136	174	30276	13.191
125	15625	11.180	175	30625	13.229
126	15876	11.225	176	30976	13.266
127	16129	11.269	177	31329	13.304
128	16384	11.314	178	31684	13.342
129	16641	11.358	179	32041	13.379
130	16900	11.402	180	32400	13.416
131	17161	11.446	181	32761	13.454
132	17424	11.489	182	33124	13.491
133	17689	11.533	183	33489	13.528
134	17956	11.576	184	33856	13.565
135	18225	11.619	185	34225	13.601
136	18496	11.662	186	34596	13.638
137	18769	11.705	187	34969	13.675
138	19044	11.747	188	35344	13.711
139	19321	11.790	189	35721	13.748
140	19600	11.832	190	36100	13.784
141	19881	11.874	191	36481	13.820
142	20164	11.916	192	36864	13.856
143	20449	11.958	193	37249	13.892
144	20736	12.000	194	37636	13.928
145	21025	12.042	195	38025	13.964
146	21316	12.083	196	38416	14.000
147	21609	12.124	197	38809	14.036
148	21904	12.166	198	39204	14.071
149	22201	12.207	199	39601	14.107
150	22500	12.247	200	40000	14.142

Table of Squares and Square Roots (Continued)

Number	Square	Square root	Number	Square	Square root
201	40401	14.177	251	63001	15.843
202	40804	14.213	252	63504	15.875
203	41209	14.248	253	64009	15.906
204	41616	14.283	254	64516	15.937
205	42025	14.318	255	65025	15.969
206	42436	14.353	256	65536	16.000
207	42849	14.387	257	66049	16.031
208	43264	14.422	258	66564	16.062
209	43681	14.457	259	67081	16.093
210	44100	14.491	260	67600	16.125
211	44521	14.526	261	68121	16.155
212	44944	14.560	262	68644	16.186
213	45369	14.595	263	69169	16.217
214	45796	14.629	264	69696	16.248
215	46225	14.663	265	70225	16.279
216	46656	14.697	266	70756	16.310
217	47089	14.731	267	71289	16.340
218	47524	14.765	268	71824	16.371
219	47961	14.799	269	72361	16.401
220	48400	14.832	270	72900	16.432
221	48841	14.866	271	73441	16.462
222	49284	14.900	272	73984	16.492
223	49729	14.933	273	74529	16.523
224	50176	14.967	274	75076	16.553
225	50625	15.000	275	75625	16.583
226	51076	15.033	276	76176	16.613
227	51529	15.067	277	76729	16.643
228	51984	15.100	278	77284	16.673
229	52441	15.133	279	77841	16.703
230	52900	15.166	280	78400	16.733
231	53361	15.199	281	78961	16.763
232	53824	15.232	282	79524	16.793
233	54289	15.264	283	80089	16.823
234	54756	15.297	284	80656	16.852
235	55225	15.330	285	81225	16.882
236	55696	15.362	286	81796	16.912
237	56169	15.395	287	82369	16.941
238	56644	15.427	288	82944	16.971
239	57121	15.460	289	83521	17.000
240	57600	15.492	290	84100	17.029
241	58081	15.524	291	84681	17.059
242	58564	15.556	292	85264	17.088
243	59049	15.588	293	85849	17.117
244	59536	15.620	294	86436	17.146
245	60025	15.652	295	87025	17.176
246	60516	15.684	296	87616	17.205
247	61009	15.716	297	88209	17.234
248	61504	15.748	298	88804	17.263
249	62001	15.780	299	89401	17.292
250	62500	15.811	300	90000	17.321

Table of Squares and Square Roots (Continued)

Number	Square	Square root	Number	Square	Square root
301	90601	17.349	351	123201	18.735
302	91204	17.378	352	123904	18.762
303	91809	17.407	353	124609	18.788
304	92416	17.436	354	125316	18.815
305	93025	17.464	355	120625	18.841
306	93636	17.493	356	126736	18.868
307	94249	17.521	357	127449	18.894
308	94864	17.550	358	128164	18.921
309	95481	17.578	359	128881	18.947
310	96100	17.607	360	129600	18.974
311	96721	17.635	361	130321	19.000
312	97344	17.664	362	131044	19.026
313	97969	17.692	363	131769	19.053
314	98596	17.720	364	132496	19.079
315	99225	17.748	365	133225	19.105
316	99856	17.776	366	133956	19.131
317	100489	17.804	367	134689	19.157
318	101124	17.833	368	135424	19.183
319	101761	17.861	369	136161	19.209
320	102400	17.889	370	136900	19.235
321	103041	17.916	371	137641	19.261
322	103684	17.944	372	138384	19.287
323	104329	17.972	373	139129	19.313
324	104976	18.000	374	139876	19.339
325	105625	18.028	375	140625	19.363
326	106276	18.055	376	141376	19.391
327	106929	18.083	377	142129	19.416
328	107584	18.111	378	142884	19.442
329	108241	18.138	379	143641	19.468
330	108900	18.166	380	144400	19.494
331	109561	18.193	381	145161	19.519
332	110224	18.221	382	145924	19.545
333	110889	18.248	383	146689	19.570
334	111556	18.276	384	147456	19.596
335	112225	18.303	385	148225	19.621
336	112896	18.330	386	148996	19.647
337	113569	18.358	387	149769	19.672
338	114244	18.385	388	150544	19.698
339	114921	18.412	389	151321	19.723
340	115600	18.439	390	152100	19.748
341	116281	18.466	391	152881	19.774
342	116964	18.493	392	153664	19.799
343	117649	18.520	393	154449	19.824
344	118336	18.547	394	155236	19.849
345	119025	18.574	395	156025	19.875
346	119716	18.601	396	156816	19.900
347	120409	18.628	397	157609	19.925
348	121104	18.655	398	158404	19.950
349	121801	18.682	399	159201	19.975
350	122500	18.708	400	160000	20.000

Table of Squares and Square Roots (Continued)

Number	Square	Square root	Number	Square	Square root
401	160801	20.025	451	203401	21.237
402	161604	20.050	452	204304	21.260
403	162409	20.075	453	205209	21.284
404	163216	20.100	454	206116	21.307
405	164025	20.125	455	207025	21.331
406	164836	20.149	456	207936	21.354
407	165649	20.174	457	208849	21.378
408	166464	20.199	458	209764	21.401
409	167281	20.224	459	210681	21.424
410	168100	20.248	460	211600	21.448
411	168921	20.273	461	212521	21.471
412	169744	20.298	462	213444	21.494
413	170569	20.322	463	214369	21.517
414	171396	20.347	464	215296	21.541
415	172225	20.372	465	216225	21.564
416	173056	20.396	466	217156	21.587
417	173889	20.421	467	218089	21.610
418	174724	20.445	468	219024	21.633
419	175561	20.469	469	219961	21.656
420	176400	20.494	470	220900	21.679
421	177241	20.518	471	221841	21.703
422	178084	20.543	472	222784	21.726
423	178929	20.567	473	223729	21.749
424	179776	20.591	474	224676	21.772
425	180625	20.616	475	225625	21.794
426	181476	20.640	476	226576	21.817
427	182329	20.664	477	227529	21.840
428	183184	20.688	478	228484	21.863
429	184041	20.712	479	229441	21.886
430	184900	20.736	480	230400	21.909
431	185761	20.761	481	231361	21.932
432	186624	20.785	482	232324	21.954
433	187489	20.809	483	233289	21.977
434	188356	20.833	484	234256	22.000
435	189225	20.857	485	235225	22.023
436	190096	20.881	486	236196	22.045
437	190969	20.905	487	237169	22.068
438	191844	20.928	488	238144	22.091
439	192721	20.952	489	239121	22.113
440	193600	20.976	490	240100	22.136
441	194481	21.000	491	241081	22.159
442	195364	21.024	492	242064	22.181
443	196249	21.048	493	243049	22.204
444	197136	21.071	494	244036	22.226
445	198025	21.095	495	245025	22.249
446	198916	21.119	496	246016	22.271
447	199809	21.142	497	247009	22.293
448	200704	21.166	498	248004	22.316
449	201601	21.190	499	249003	22.338
450	202500	21.213	500	250000	22.361

Table of Squares and Square Roots (Continued)

Number	Square	Square root	Number	Square	Square root
501	251001	22.383	551	303601	23.473
502	252004	22.405	552	304704	23.495
503	253009	22.428	553	305809	23.516
504	254016	22.450	554	306916	23.537
505	255025	22.472	555	308025	23.558
506	256036	22.494	556	309136	23.580
507	257049	22.517	557	310249	23.601
508	258064	22.539	558	311364	23.622
509	259081	22.561	559	312481	23.643
510	260100	22.583	560	313600	23.664
511	261121	22.605	561	314721	23.685
512	262144	22.627	562	315844	23.707
513	263169	22.650	563	316969	23.728
514	264196	22.672	564	318096	23.749
515	265225	22.694	565	319225	23.770
516	266256	22.716	566	320356	23.791
517	267289	22.738	567	321489	23.812
518	268324	22.760	568	322624	23.833
519	269361	22.782	569	323761	23.854
520	270400	22.804	570	324900	23.875
521	271441	22.825	571	326041	23.896
522	272484	22.847	572	327184	23.917
523	273529	22.869	573	328329	23.937
524	274576	22.891	574	329476	23.958
525	275625	22.913	575	330625	23.979
526	276676	22.935	576	331776	24.000
527	277729	22.956	577	332929	24.021
528	278784	22.978	578	334084	24.042
529	279841	23.000	579	335241	24.062
530	280900	23.022	580	336400	24.083
531	281961	23.043	581	337561	24.104
532	283024	23.065	582	338724	24.125
533	284089	23.087	583	339889	24.145
534	285156	23.108	584	341056	24.166
535	286225	23.130	585	342225	24.187
536	287296	23.152	586	343396	24.207
537	288369	23.173	587	344569	24.228
538	289444	23.195	588	345744	24.249
539	290521	23.216	589	346921	24.269
540	291600	23.238	590	348100	24.290
541	292681	23.259	591	349281	24.310
542	293764	23.281	592	350464	24.331
543	294849	23.302	593	351649	24.352
544	295936	23.324	594	352836	24.372
545	297025	23.345	595	354025	24.393
546	298116	23.367	596	355216	24.413
547	299209	23.388	597	356409	24.434
548	300304	23.409	598	357604	24.454
549	301401	23.431	599	358801	24.474
550	302500	23.452	600	360000	24.495

Table of Squares and Square Roots (Continued)

Number	Square	Square root	Number	Square	Square root
601	361201	24.515	651	423801	25.515
602	362404	24.536	652	425104	25.534
603	363609	24.556	653	426409	25.554
604	364816	24.576	654	427716	25.573
605	366025	24.597	655	429025	25.593
606	367236	24.617	656	430336	25.612
607	368449	24.637	657	431649	25.632
608	369664	24.658	658	432964	25.652
609	370881	24.678	659	434281	25.671
610	372100	24.698	660	435600	25.690
611	373321	24.718	661	436921	25.710
612	374544	24.739	662	438244	25.729
613	375769	24.759	663	439569	25.749
614	376996	24.779	664	440896	25.768
615	378225	24.799	665	442225	25.788
616	379456	24.819	666	443556	25.807
617	380689	24.839	667	444889	25.826
618	381924	24.860	668	446224	25.846
619	383161	24.880	669	447561	25.865
620	384400	24.900	670	448900	25.884
621	385641	24.920	671	450241	25.904
622	386884	24.940	672	451584	25.923
623	388129	24.960	673	452929	25.942
624	389376	24.980	674	454276	25.962
625	390625	25.000	675	455625	25.981
626	391876	25.020	676	456976	26.000
627	393129	25.040	677	458329	26.019
628	394384	25.060	678	459684	26.038
629	395641	25.080	679	461041	26.058
630	396900	25.100	680	462400	26.077
631	398161	25.120	681	463761	26.096
632	399424	25.140	682	465124	26.115
633	400689	25.159	683	466489	26.134
634	401956	25.179	684	467856	26.153
635	403225	25.199	685	469225	26.173
636	404496	25.219	686	470596	26.192
637	405769	25.239	687	471969	26.211
638	407044	25.259	688	473344	26.230
639	408321	25.278	689	474721	26.249
640	409600	25.298	690	476100	26.268
641	410881	25.318	691	477481	26.287
642	412164	25.338	692	478864	26.306
643	413449	25.357	693	480249	26.325
644	414736	25.377	694	481636	26.344
645	416025	25.397	695	483025	26.363
646	417316	25.417	696	484416	26.382
647	418609	25.436	697	485809	26.401
648	419904	25.456	698	487204	26.420
649	421201	25.475	699	488601	26.439
650	422500	25.495	700	490000	26.458

Table of Squares and Square Roots (Continued)

Number	Square	Square root	Number	Square	Square root
701	491401	26.476	751	564001	27.404
702	492804	26.495	752	565504	27.423
703	404209	26.514	753	567009	27.441
704	495616	26.533	754	568516	27.459
705	497025	26.552	755	570025	27.477
706	498436	26.571	756	571536	27.495
707	499849	26.589	757	573049	27.514
708	501264	26.608	758	574564	27.532
709	502681	26.627	759	576081	27.550
710	504100	26.646	760	577600	27.568
711	505521	26.665	761	579121	27.586
712	506944	26.683	762	580644	27.604
713	508369	26.702	763	582169	27.622
714	509796	26.721	764	583696	27.641
715	511225	26.739	765	585225	27.659
716	512656	26.758	766	586756	27.677
717	514089	26.777	767	588289	27.695
718	515524	26.796	768	589824	27.713
719	516961	26.814	769	591361	27.731
720	518400	26.833	770	592900	27.749
721	519841	26.851	771	594441	27.767
722	521284	26.870	772	595984	27.785
723	522729	26.889	773	597529	27.803
724	524176	26.907	774	599076	27.821
725	525625	26.926	775	600625	27.839
726	527076	26.944	776	602176	27.857
727	528529	26.963	777	603729	27.875
728	529984	26.981	778	605284	27.893
729	531441	27.000	779	606841	27.911
730	532900	27.019	780	608400	27.928
731	534361	27.037	781	609961	27.946
732	534824	27.055	782	611524	27.964
733	537289	27.074	783	613089	27.982
734	538756	27.092	784	614656	28.000
735	540225	27.111	785	616225	28.018
736	541696	27.129	786	617796	28.036
737	543169	27.148	787	619369	28.054
738	544644	27.166	788	620944	28.071
739	546121	27.185	789	622521	28.089
740	547600	27.203	790	624100	28.107
741	549081	27.221	791	625681	28.125
742	550564	27.240	792	627264	28.142
743	552049	27.258	793	628849	28.160
744	553536	27.276	794	630436	28.178
745	555025	27.295	795	632025	28.196
746	556516	27.313	796	633616	28.213
747	558009	27.331	797	635209	28.231
748	559504	27.350	798	636804	28.249
749	561001	27.368	799	638401	28.267
750	562500	27.386	800	640000	28.284

Table of Squares and Square Roots (Continued)

Number	Square	Square root
801	641601	28.302
802	643204	28.320
803	644809	28.337
804	646416	28.355
805	648025	28.373
806	649636	28.390
807	651249	28.408
808	652864	28.425
809	654481	28.443
810	656100	28.460
811	657721	28.478
812	659344	28.496
813	660969	28.513
814	662596	28.531
815	664225	28.548
816	665856	28.566
817	667489	28.583
818	669124	28.601
819	670761	28.618
820	672400	28.636
821	674041	28.653
822	675684	28.671
823	677329	28.688
824	678976	28.705
825	680625	28.723
826	682276	28.740
827	683929	28.758
828	685584	28.775
829	687241	28.792
830	688900	28.810
831	690561	28.827
832	692224	28.844
833	693889	28.862
834	695556	28.879
835	697225	28.896
836	698896	28.914
837	700569	28.931
838	702244	28.948
839	703921	28.965
840	705600	28.983
841	707281	29.000
842	708964	29.017
843	710649	29.034
844	712336	29.052
845	714025	29.069
846	715716	29.086
847	717409	29.103
848	719104	29.120
849	720801	29.138
850	722500	29.155

Number	Square	Square root
851	724201	29.172
852	725904	29.189
853	727609	29.206
854	729316	29.223
855	731025	29.240
856	732736	29.257
857	734449	29.275
858	736164	29.292
859	737881	29.309
860	739600	29.326
861	741321	29.343
862	743044	29.360
863	744769	29.377
864	746496	29.394
865	748225	29.411
866	749956	29.428
867	751689	29.445
868	753424	29.462
869	755161	29.479
870	756900	29.496
871	758641	29.513
872	760384	29.530
873	762129	29.547
874	763876	29.563
875	765625	29.580
876	767376	29.597
877	769129	29.614
878	770884	29.361
879	772641	29.648
880	774400	29.665
881	776161	29.682
882	777924	29.698
883	779689	29.715
884	781456	29.732
885	783225	29.749
886	784996	29.766
887	786769	29.783
888	788544	29.799
889	790321	29.816
890	792100	29.833
891	793881	29.850
892	795664	29.866
893	797449	29.883
894	799236	29.900
895	801025	29.916
896	802816	29.933
897	804609	29.950
898	806404	29.967
899	808201	29.983
900	810000	30.000

Table of Squares and Square Roots (Continued)

Number	Square	Square root	Number	Square	Square root
901	811801	30.017	951	904401	30.838
902	813604	30.033	952	906304	30.854
903	815409	30.050	953	908209	30.871
904	817216	30.067	954	910116	30.887
905	819025	30.083	955	912025	30.903
906	820836	30.100	956	913936	30.919
907	822649	30.116	957	915849	30.935
908	824464	30.133	958	917764	30.952
909	826281	30.150	959	919681	30.968
910	828100	30.166	960	921600	30.984
911	829921	30.183	961	923521	31.000
912	831744	30.199	962	925444	31.016
913	833569	30.216	963	927369	31.032
914	835396	30.232	964	929296	31.048
915	837225	30.249	965	931225	31.064
916	839056	30.265	966	933156	31.081
917	840889	30.282	967	935089	31.097
918	842724	30.299	968	937024	31.113
919	844561	30.315	969	938961	31.129
920	846400	30.332	970	940900	31.145
921	848241	30.348	971	942841	31.161
922	850084	30.364	972	944784	31.177
923	851929	30.381	973	946729	31.193
924	853776	30.397	974	948676	31.209
925	855625	30.414	975	950625	31.225
926	857476	30.430	976	952576	31.241
927	859329	30.447	977	954529	31.257
928	861184	30.463	978	956484	31.273
929	863041	30.480	979	958441	31.289
930	864900	30.496	980	960400	31.305
931	866761	30.512	981	962361	31.321
932	868624	30.529	982	964324	31.337
933	870489	30.545	983	966289	31.353
934	872356	30.561	984	968256	31.369
935	874225	30.578	985	970225	31.385
936	876096	30.594	986	972196	31.401
937	877969	30.610	987	974169	31.417
938	879844	30.627	988	976144	31.432
939	881721	30.643	989	978121	31.448
940	883600	30.659	990	980100	31.464
941	885481	30.676	991	982081	31.480
942	887364	30.692	992	984064	31.496
943	889249	30.708	993	986049	31.512
944	891136	30.725	994	988036	31.528
945	893025	30.741	995	990025	31.544
946	894916	30.757	996	992016	31.559
947	896809	30.773	997	994009	31.575
948	898704	30.790	998	996004	31.591
949	900601	30.806	999	998001	31.607
950	902500	30.822	1000	1000000	31.623

EXAMPLES OF VARIABILITY FOUND IN THE MEAT INDUSTRY

Examples of control limits and standard deivations used in the meat industry.

<u>Weight</u>

1000 lbs scale wts (standard weight used)	0.4% deviation or 4 lbs.
Pallet test (components equal total)	0.4% deviation
30-36 lbs meat units (weight)	0.5 lbs deviation
30 lbs. box beef patties	±60 g
30 lbs. boxes of beef patties	σ20 g
Whole chicken	σ0.44 lbs.
32 oz. package of chicken	σ1.4 oz.
2 lbs. frankfurter package	σ0.070 lbs.
16 oz. package of frankfurters	±1/8 oz. Target + 1/8 oz. UCL at 2/8 oz. LCL at 0
1 lb. wiener package	σ0.11 lbs.
1 lb. (10 to 1) sliced bologna	σ0.064 lbs.
16 oz. bacon package weight	$\sigma$of 0.25 oz.
6/1 lbs. beef patties	σ3.2 g
3/1 lbs. hamburger pattie	σ0.014 lbs.
4/1 lbs. beef patties	σ0.45 oz.
3/1 lbs. round steaks	σ7.6 g to 0.03 oz.
4/1 lbs. beef round steaks	σ0.26 oz.
5 oz. pork chops with inexperienced cutter	σ0.6 oz.
5 oz. cube steaks	σ0.34 oz.
12 oz. strip steaks	σ0.0125 lbs.
One frankfurter	σ2.9 g to 6/32 oz.
30 g smokies	σ1.5 g
Bacon slices	σ3.9 g
Number of pieces of pepper or pimento in P&P slice ($\bar{X}$=28 g)	σ0.11
Hot carcass weight of slaughter weight hogs	σ10.75 lbs.
Moisture composition of pork muscle	σ1.75 lbs.
Ham thickness in cm of slaughter weight hogs	σ0.80 lbs.
One pound packages of hamburger	σ0.10 lbs.
Fresh pork sausage links, 2 oz.	σ0.35 g

WEIGHT LIMITS OFTEN USED IN THE MEAT AREA

	Bacon	5 oz. Slice pack	8 oz. Slice pack	16 oz. Slice pack
No package out of 10 above		+3/8 oz.		
No more than one package out of 10 above	+1/2 oz.		1/4 oz.	
No more than two packages out of 10 above				+3/8 oz.
No package out of 10 less than	-1/4 oz.	-1/4 oz.		
No more than one package out of 10 less than			-1/4 oz.	
No more than two packages out of 10 less than				-1/4 oz.
Scalers accept	-1/4 to + 3/8 oz.	-1/8 oz. + 1/4 oz.		
Limits on total of 10 packages	0 to + 1-1/4 oz.	0 to + 1-1/4 oz.	0 to + 1-1/2 oz.	0 to + 1-3/4 oz.

EXAMPLES OF VARIABILITY IN PERCENTAGE COMPOSITION OF MEAT PRODUCTS

Anylray vs ether extract	±0.4 to 1%
Fat percentage of ground meat	±2% from desired
90% lean beef	±2% fat
10/1 pound beef patties	16-19% fat
1/4 lbs. pattie beef	20-23% fat
Reg. ground beef fat content	σ2.24% fat
Percentage fat in comminuted product	σ0.53% fat
Regular grind meat (24% fat)	σ3.6% fat
Chopped sirloin (18% fat)	σ3.1% fat
Percent fat in emulsion cooked product that was preblended	σ0.59% fat
Percent water in emulsion cooked product that was preblended	σ0.62% water
Percent protein in emulsion cooked product that was preblended	σ0.18% protein
Percent added water in emulsion cooked product that was preblended	σ0.81% added water
Percentage white in bacon window	Max. 25%

Examples of variability found in miscellaneous meat items.

Vacuum package leakers	6% leakers
Vacuum package bacon leakers	0-11% range 0-2% specifications
Smokehouse shrinkage (franks)	2.7-4%
Franks smokehouse shrink	2.8-3.7%

A Two Sample Tolerance Test

A good management tool or can be used to examine an inspector's results.

The two sample tolerance test indicates the difference allowed in findings between two different samples taken from the same lot. The tables indicate the amount of difference you can expect due to sampling variation alone. If the samples differ by more than the table values it suggests that something in addition to sample variation is causing these differences.

ATTRIBUTE (Discrete) sample use Table I (10% chance of being wrong) when both samples are the same size.

1. Samples must be the same size
2. Enter column one or four with the number of defects in the first sample.
3. Upper and lower limits for two samples are listed in column 2 and 3 or in columns 5 and 6.
4. Example: 50 boxes were examined and 8 were found to contain defects in sample 1. Fifty boxes were examined in sample 2 and sample variation indicates they should contain between 3 and 16 defects. If the second samples contains 2 or less or 17 or more, then something other than normal variability 9 times out of 10 caused this much difference.

ATTRIBUTE (Discrete) sample use Table II (10% chance of being wrong) when the two samples are of different sizes.

1. Determine ratio of sample size by placing the smaller sample in the numerator and the larger sample in the denominator.
2. Enter Table II in the closest ratio (columns 2 through 11 labelled 1 through 1/10).

3. Enter table II in column one with the number of defects in the larger sample.
4. The upper and lower limits of the second sample are shown under the appropriate sample size ratio.
5. Example:

	Sample Size	Defects
Sample 1	500	24
Sample 2	100	12

Would have expected sample 2 to containing between 2 and 9 defects and since it contained 12 probably something is involved other than normal variation.

Variable (continuous) samples. Use Table III (10% chance of being wrong)

1. Find the mean and range for each of 2 samples.
2. Calculate difference between 2 sample means

$$d = \bar{X}_1 - \bar{X}_2$$

average range of 2 samples

$$\bar{R} = \frac{R_1 + R_2}{2}$$

3. limit

$$\frac{d}{\bar{R}} = \frac{\bar{X}_1 - X_2}{R_1 + R_{2/2}}$$

4. Compare to table value at appropriate sample size. If it does not exceed the limit, then samples agree.

5. Example: 16 oz. package of meat to nearest 0.1 oz.

	Sample 1	Sample 2
	16.1	16.1
	16.2	16.1
	15.9	16.1
	15.8	16.2
	16.2	16.1
ΣX	80.2	80.6
$\bar{X}$	16.04	16.12
R	0.4	0.1

$$d = 16.12 - 16.04 = 0.08$$

$$\bar{R} = \frac{0.4 + 0.1}{2} = 0.25$$

$$\frac{d}{\bar{R}} = \frac{0.08}{0.25} = 0.32$$

Since $d/\bar{R}$ is less than 0.050 for sample size of 5, the 2 samples agree

Table I

TWO SAMPLE COMPARISON (EQUAL SAMPLE SIZES)
LIMITS FOR ATTRIBUTES TYPE SAMPLES

10% chance of being wrong

NO. OF DEFECTS (DEFECTIVES) IN FIRST SAMPLE	LIMITS FOR SECOND SAMPLE LOWER	UPPER	NO. OF DEFECTS (DEFECTIVES) IN FIRST SAMPLE	LIMITS FOR SECOND SAMPLE LOWER	UPPER
0	0	2	51	36	68
1	0	4	52	37	70
2	0	6	53	38	71
3	1	8	54	39	72
4	1	10	55	40	73
5	2	11	56	40	74
6	2	13	57	41	75
7	3	14	58	42	77
8	3	16	59	43	78
9	4	17	60	44	79
10	4	18	61	45	80
11	5	20	62	46	81
12	6	21	63	46	82
13	6	22	64	47	83
14	7	24	65	48	85
15	8	25	66	49	86
16	8	26	67	50	87
17	9	27	68	51	88
18	10	29	69	52	89
19	11	30	70	52	90
20	11	31	71	53	91
21	12	33	72	54	93
22	13	34	73	55	94
23	14	35	74	56	95
24	14	36	75	57	96
25	15	38	76	58	97
26	16	39	77	58	98
27	17	40	78	59	99
28	18	41	79	60	101
29	18	42	80	61	102
30	19	44	81	62	103
31	20	45	82	63	104
32	21	46	83	64	105
33	21	47	84	65	106
34	22	48	85	65	107
35	23	50	86	66	108
36	24	51	87	67	110
37	25	52	88	68	111
38	25	53	89	69	112
39	26	54	90	70	113
40	27	56	91	71	114
41	28	57	92	72	115
42	29	58	93	72	116
43	30	59	94	73	117
44	30	60	95	74	119
45	31	61	96	75	120
46	32	63	97	76	121
47	33	64	98	77	122
48	34	65	99	78	123
49	35	66	100	79	124
50	35	67			

TABLE II (10% Chance of being wrong)

TWO SAMPLE COMPARISON (UNEQUAL SIZE SAMPLE)
LIMITS FOR ATTRIBUTE TYPE SAMPLES

10% chance of being wrong

ACCEPTANCE NUMBERS FOR SMALLER SAMPLE

Ratio (R) = $\frac{\text{Smaller sample size}}{\text{Larger sample size}}$

NUMBER DEFECTIVES (DEFECTS) IN LARGER SAMPLE	R = 1		R = 1/2		R = 1/3		R = 1/4		R = 1/5		R = 1/6		R = 1/7		R = 1/8		R = 1/9		R = 1/10	
	L	U	L	U	L	U	L	U	L	U	L	U	L	U	L	U	L	U	L	U
0	0	2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	4	0	2	0	2	0	1	0	1	0	1	0	1	0	1	0	1	0	0
2	0	6	0	3	0	2	0	2	0	2	0	1	0	1	0	1	0	1	0	1
3	1	8	0	4	0	3	0	2	0	2	0	2	0	1	0	1	0	1	0	1
4	1	10	0	5	0	4	0	3	0	2	0	2	0	2	0	2	0	1	0	1
5	2	11	1	6	0	4	0	3	0	3	0	2	0	2	0	2	0	2	0	2
6	2	13	1	7	0	5	0	4	0	3	0	3	0	2	0	2	0	2	0	2
7	3	14	1	8	1	5	0	4	0	4	0	3	0	3	0	2	0	2	0	2
8	3	16	1	8	1	6	0	5	0	4	0	3	0	3	0	3	0	2	0	2
9	4	17	2	9	1	6	1	5	0	4	0	4	0	3	0	3	0	3	0	2
10	4	18	2	10	1	7	1	5	0	5	0	4	0	3	0	3	0	3	0	3
11	5	20	2	11	1	7	1	6	1	5	0	4	0	4	0	3	0	3	0	3
12	6	21	2	11	1	8	1	6	1	5	1	4	0	4	0	4	0	3	0	3
13	6	22	3	12	2	8	1	7	1	6	1	5	0	4	0	4	0	3	0	3
14	7	24	3	13	2	9	1	7	1	6	1	5	1	4	0	4	0	4	0	3
15	8	25	3	13	2	9	1	7	1	6	1	5	1	5	0	4	0	4	0	3
16	8	26	4	14	2	10	1	8	1	6	1	6	1	5	1	4	0	4	0	4
17	9	27	4	15	2	10	2	8	1	7	1	6	1	5	1	5	0	4	0	4
18	10	29	4	15	3	11	2	8	1	7	1	6	1	5	1	5	1	4	0	4
19	11	30	5	16	3	11	2	9	1	7	1	6	1	6	1	5	1	4	0	4
20	11	31	5	17	3	12	2	9	1	8	1	7	1	6	1	5	1	5	1	4
21	12	33	5	17	3	12	2	10	2	8	1	7	1	6	1	5	1	5	1	4
22	13	34	6	18	3	13	2	10	2	8	1	7	1	6	1	6	1	5	1	5
23	14	35	6	19	4	13	2	10	2	8	1	7	1	6	1	6	1	5	1	5
24	14	36	6	19	4	14	3	11	2	9	1	7	1	7	1	6	1	5	1	5
25	15	38	7	20	4	14	3	11	2	9	2	8	1	7	1	6	1	6	1	5
26	16	39	7	21	4	14	3	11	2	9	2	8	1	7	1	6	1	6	1	5
27	17	40	7	21	4	15	3	12	2	10	2	8	1	7	1	7	1	6	1	5
28	18	41	8	22	5	15	3	12	2	10	2	8	1	7	1	7	1	6	1	6
29	18	42	8	23	5	16	3	12	2	10	2	9	2	8	1	7	1	6	1	6
30	19	44	9	23	5	16	4	13	3	10	2	9	2	8	1	7	1	6	1	6
31	20	45	9	24	5	17	4	13	3	11	2	9	2	8	1	7	1	7	1	6
32	21	46	9	24	6	17	4	13	3	11	2	9	2	8	1	7	1	7	1	6
33	21	47	10	25	6	17	4	14	3	11	2	10	2	8	2	8	1	7	1	6
34	22	48	10	26	6	18	4	14	3	12	2	10	2	9	2	8	1	7	1	6
35	23	50	10	26	6	18	4	14	3	12	3	10	2	9	2	8	1	7	1	7

same as in Table 1

TABLE III.

TWO SAMPLE COMPARISON (EQUAL SAMPLE SIZES)
LIMITS FOR VARIABLE TYPE OF SAMPLES

10% Chance of being wrong

Subgroup Sample Size	Limit for $d/\overline{R}$
3	1.00
5	0.50
10	0.25

Where d = difference between two sample averages

$\overline{R}$ = average range of the two samples,

$$\overline{R} = \frac{R_1 + R_2}{2}$$

IBM Card Box Key

Set #1 (some variation, different means)

Box Code	Mean	Standard Deviation
M	16.82	0.307
L	16.61	0.307
K	16.40	0.307
A	16.20	0.305
I	16.00	0.307
B & H	15.80	0.307
G	15.62	0.307
F	15.42	0.307
E	15.22	0.307

Set #2 (similar mean, different variation)

D	16.19	0.560
J	16.21	0.307
C	16.07	0.080

GROUP A NET WEIGHTS PROC FOODS

N	505	RANGE	2.39999	COEF. VAR.	0.01878
MEAN	16.21802	VARIANCE	0.09275	SKEWNESS	0.00000
MEDIAN	16.20291	STD. DEV.	0.30455	NO. OF CELLS	25

```
CELL MID.     FREQ.
  14.9000        1  *
  15.0000        0
  15.1000        0
  15.2000        0
  15.3000        0
  15.4000        2  **
  15.5000        2  **
  15.6000        3  ***
  15.7000       14  **************
  15.8000       30  ******************************
  15.9000       49  *************************************************
  16.0000       50  **************************************************
  16.1000       65  *****************************************************************
  16.2000       68  ********************************************************************
  16.3000       56  ********************************************************
  16.4000       49  *************************************************
  16.5000       40  ****************************************
  16.6000       37  *************************************
  16.7000       20  ********************
  16.8000        9  *********
  16.9000        6  ******
  17.0000        2  **
  17.1000        1  *
  17.2000        0
  17.3000        1  *
```

GROUP B & H NET WEIGHTS PROC FOODS

N	507	RANGE	2.39999	COEF. VAR.	0.01943
MEAN	15.81834	VARIANCE	0.09442	SKEWNESS	0.00000
MEDIAN	15.80440	STD. DEV.	0.30729	NO. OF CELLS	25

CELL MID.	FREQ.	
14.5000	1	*
14.6000	0	
14.7000	0	
14.8000	0	
14.9000	2	**
15.0000	1	*
15.1000	1	*
15.2000	3	***
15.3000	14	**************
15.4000	31	*******************************
15.5000	49	***
15.6000	50	**
15.7000	64	**
15.8000	68	**
15.9000	56	**
16.0000	50	**
16.1000	40	**
16.2000	37	*************************************
16.3000	20	********************
16.4000	10	**********
16.5000	6	******
16.5999	2	**
16.6999	1	*
16.7999	0	
16.8999	1	*

GROUP C NET WEIGHTS PROC FOODS

```
N          492          RANGE        0.44999    COEF. VAR.      0.00499
MEAN        16.07520    VARIANCE     0.00644    SKEWNESS        0.00000
MEDIAN      16.07497    STD. DEV.    0.08025    NO. OF CELLS   10

CELL MID.   FREQ.
  15.8500      3   *
  15.9000     10   *****
  15.9500     34   *****************
  16.0000     90   *********************************************
  16.0500    109   ******************************************************
  16.1000    109   ******************************************************
  16.1500     90   *********************************************
  16.2000     33   *****************
  16.2499     10   *****
  16.2999      4   **
```

GROUP D NET WEIGHTS PROC FOODS

```
N          505          RANGE        3.19999    COEF. VAR.      0.03458
MEAN        16.19901    VARIANCE     0.31387    SKEWNESS        0.00000
MEDIAN      16.19664    STD. DEV.    0.56024    NO. OF CELLS   33

CELL MID.   FREQ.
  14.6000      1   *
  14.7000      2   **
  14.8000      2   **
  14.9000      3   ***
  15.0000      3   ***
  15.1000      4   ****
  15.2000     10   **********
  15.3000     10   **********
  15.4000     15   ***************
  15.5000     15   ***************
  15.6000     15   ***************
  15.7000     20   ********************
  15.8000     25   *************************
  15.9000     30   ******************************
  16.0000     35   ***********************************
  16.1000     41   *****************************************
  16.2000     45   *********************************************
  16.3000     40   ****************************************
  16.4000     36   ************************************
  16.5000     28   ****************************
  16.5999     25   *************************
  16.6999     20   ********************
  16.7999     15   ***************
  16.8999     15   ***************
  16.9999     15   ***************
  17.0999     10   **********
  17.1999     10   **********
  17.2999      4   ****
  17.3999      3   ***
  17.4999      3   ***
  17.5999      2   **
  17.6998      2   **
  17.7998      1   *
```

GROUP E NET WEIGHTS PROC FOODS

```
N            507          RANGE        2.39999     COEF. VAR.      0.02019
MEAN          15.21834    VARIANCE     0.09443     SKEWNESS        1.00080
MEDIAN        15.20440    STD. DEV.    0.30729     NO. OF CELLS   25

CELL MID.        FREQ.
  13.9000            1  *
  14.0000            0
  14.1000            0
  14.2000            0
  14.3000            2  **
  14.4000            1  *
  14.5000            1  *
  14.6000            3  ***
  14.7000           14  **************
  14.8000           31  *******************************
  14.9000           49  *************************************************
  15.0000           50  **************************************************
  15.1000           64  ****************************************************************
  15.2000           68  ********************************************************************
  15.3000           56  ********************************************************
  15.4000           50  **************************************************
  15.5000           40  ****************************************
  15.6000           37  *************************************
  15.7000           20  ********************
  15.8000           10  **********
  15.9000            6  ******
  16.0000            2  **
  16.1000            1  *
  16.2000            0
  16.3000            1  *
```

GROUP F NET WEIGHTS PROC FOODS

```
N          507         RANGE        2.40000    COEF. VAR.      0.01993
MEAN        15.41834   VARIANCE     0.09443    SKEWNESS        1.00079
MEDIAN      15.40440   STD. DEV.    0.30729    NO. OF CELLS   25

CELL MID.      FREQ.
  14.1000          1  *
  14.2000          0
  14.3000          0
  14.4000          0
  14.5000          2  **
  14.6000          1  *
  14.7000          1  *
  14.8000          3  ***
  14.9000         14  **************
  15.0000         31  *******************************
  15.1000         49  *************************************************
  15.2000         50  **************************************************
  15.3000         65  *****************************************************************
  15.4000         68  ********************************************************************
  15.5000         56  ********************************************************
  15.6000         50  **************************************************
  15.7000         40  ****************************************
  15.8000         37  *************************************
  15.9000         20  ********************
  16.0000         10  **********
  16.1000          6  ******
  16.2000          2  **
  16.3000          1  *
  16.3999          0
  16.4999          1  *
```

```
                         GROUP G NET WEIGHTS PROC FOODS

N               507           RANGE          2.40000      COEF. VAR.        0.01967
MEAN             15.61934     VARIANCE       0.09443      SKEWNESS          1.00077
MEDIAN           15.60440     STD. DEV.      0.30729      NO. OF CELLS     25

CELL MID.           FREQ.
  14.3000              1  *
  14.4000              0
  14.5000              0
  14.6000              0
  14.7000              2  **
  14.8000              1  *
  14.9000              1  *
  15.0000              3  ***
  15.1000             14  **************
  15.2000             31  *******************************
  15.3000             49  *************************************************
  15.4000             50  **************************************************
  15.5000             65  *****************************************************************
  15.6000             68  ********************************************************************
  15.7000             56  ********************************************************
  15.8000             50  **************************************************
  15.9000             40  ****************************************
  16.0000             37  *************************************
  16.1000             20  ********************
  16.2000             10  **********
  16.3000              6  ******
  16.3999              2  **
  16.4999              1  *
  16.5999              0
  16.6999              1  *
```

GROUP I NET WEIGHTS PROC FOODS

```
N              508          RANGE         2.39999     COEF. VAR.          0.01916
MEAN            16.01830    VARIANCE      0.09424     SKEWNESS            1.00073
MEDIAN          16.00505    STD. DEV.     0.30698     NO. OF CELLS       25

CELL MID.          FREQ.
  14.7000              1  *
  14.8000              0
  14.9000              0
  15.0000              0
  15.1000              2  **
  15.2000              1  *
  15.3000              1  *
  15.4000              3  ***
  15.5000             14  **************
  15.6000             31  *******************************
  15.7000             49  *************************************************
  15.8000             50  **************************************************
  15.9000             65  *****************************************************************
  16.0000             69  *********************************************************************
  16.1000             56  ********************************************************
  16.2000             50  **************************************************
  16.3000             40  ****************************************
  16.4000             37  *************************************
  16.5000             20  ********************
  16.5999             10  **********
  16.6999              6  ******
  16.7999              2  **
  16.8999              1  *
  16.9999              0
  17.0999              1  *
```

```
                         GROUP J NET WEIGHTS PROC FOODS

N             507            RANGE         2.39999      COEF. VAR.        0.01892
MEAN           16.21755      VARIANCE      0.09414      SKEWNESS          1.00072
MEDIAN         16.20360      STD. DEV.     0.30682      NO. OF CELLS     25

CELL MID.          FREQ.
  14.9000              1  *
  15.0000              0
  15.1000              0
  15.2000              0
  15.3000              2  **
  15.4000              1  **
  15.5000              1  *
  15.6000              3  ***
  15.7000             14  **************
  15.8000             31  *******************************
  15.9000             49  *************************************************
  16.0000             50  **************************************************
  16.1000             65  *****************************************************************
  16.2000             69  *********************************************************************
  16.3000             56  ********************************************************
  16.4000             50  **************************************************
  16.5000             40  ****************************************
  16.5999             36  ************************************
  16.6999             20  ********************
  16.7999             10  **********
  16.8999              6  ******
  16.9999              2  **
  17.0999              1  *
  17.1999              0
  17.2999              1  *
```

```
                    GROUP K NET WEIGHTS PROC FOODS

N              507          RANGE        2.40000     COEF. VAR.        0.01872
MEAN            16.41834    VARIANCE     0.09443     SKEWNESS          1.00070
MEDIAN          16.40436    STD. DEV.    0.30729     NO. OF CELLS     25

CELL MID.          FREQ.
  15.1000              1  *
  15.2000              0
  15.3000              0
  15.4000              0
  15.5000              2  **
  15.6000              1  *
  15.7000              1  *
  15.8000              3  ***
  15.9000             14  **************
  16.0000             31  *******************************
  16.1000             49  *************************************************
  16.2000             50  **************************************************
  16.3000             65  *****************************************************************
  16.4000             68  ********************************************************************
  16.5000             56  ********************************************************
  16.5999             50  **************************************************
  16.6999             40  ****************************************
  16.7999             37  *************************************
  16.8999             20  ********************
  16.9999             10  **********
  17.0999              6  ******
  17.1999              2  **
  17.2999              1  *
  17.3999              0
  17.4999              1  *
```

GROUP L NET WEIGHTS PROC FOODS

```
N              507          RANGE        2.40000      COEF. VAR.      0.01849
MEAN            16.61834    VARIANCE     0.09443      SKEWNESS        1.00069
MEDIAN          16.60434    STD. DEV.    0.30729      NO. OF CELLS   25

CELL MID.         FREQ.
  15.3000            1  *
  15.4000            0
  15.5000            0
  15.6000            0
  16.7000            2  **
  15.8000            1  *
  15.9000            1  *
  16.0000            3  ***
  16.1000           14  **************
  16.2000           31  *******************************
  16.3000           49  *************************************************
  16.4000           50  **************************************************
  16.5000           65  *****************************************************************
  16.5999           68  ********************************************************************
  16.6999           56  ********************************************************
  16.7999           50  **************************************************
  16.8999           40  ****************************************
  16.9999           37  *************************************
  17.0999           20  ********************
  17.1999           10  **********
  17.2999            6  ******
  17.3999            2  **
  17.4999            1  *
  17.5999            0
  17.6998            1  *
```

```
                         GROUP M NET WEIGHTS PROC FOODS

N                507            RANGE           2.39999     COEF. VAR.          0.01827
MEAN              16.81834      VARIANCE        0.09443     SKEWNESS            1.00068
MEDIAN            16.80432      STD. DEV.       0.30729     NO. OF CELLS       25

CELL MID.            FREQ.
  15.5000                1  *
  15.6000                0
  15.7000                0
  15.8000                0
  15.9000                2  **
  16.0000                1  *
  16.1000                1  *
  16.2000                3  ***
  16.3000               14  **************
  16.4000               31  *******************************
  16.5000               49  *************************************************
  16.5999               50  **************************************************
  16.6999               65  *****************************************************************
  16.7999               68  ********************************************************************
  16.8999               56  ********************************************************
  16.9999               50  **************************************************
  17.0999               40  ****************************************
  17.1999               37  *************************************
  17.2999               20  ********************
  17.3999               10  **********
  17.4999                6  ******
  17.5999                2  **
  17.6998                1  *
  17.7998                0
  17.8998                1  *
```

5 SAMPLES DRAWN FROM SAMPLE A-1
BEEF PATTIES 6/1 by 50 STUDENTS

Range of the Averages	Frequency of occurrence
74.2	1
74.3	0
74.4	0
74.5	0
74.6	0
74.7	0
74.8	0
74.9	0
75.0	11
75.1	0
75.2	11
75.3	0
75.4	111
75.5	1111 11
75.6	1111 1
75.7	1111 111
75.8	111
75.9	1111 11
76.0	1
76.1	1111 1
76.2	0
76.3	0
76.4	11
76.5	1
76.6	1
3785.7	50

$\bar{X}$ 75.71

Sample A-1 Beef Patties 6/1

N	100	RANGE	4.8	COEF. VAR.	
MEAN	75.72	VARIANCE		SKEWNESS	
MEDIAN	75.75	STD. DEV.	.6599	NO. OF CELLS	38

EQUAL OR MORE THAN	FREQ.	
73.0	1	*
73.1	1	*
73.4	1	*
73.5	1	*
74.0	1	*
74.1	1	*
74.2	2	**
74.3	6	******
74.4	6	******
74.5	13	*************
74.6	7	*******
74.7	6	******
74.8	9	*********
74.9	11	***********
75.0	10	**********
75.1	16	****************
75.2	15	***************
75.3	21	*********************
75.4	19	*******************
75.5	46	**
75.6	34	**********************************
75.7	44	**
75.8	32	********************************
75.9	31	*******************************
76.0	24	************************
76.1	33	*********************************
76.2	20	********************
76.3	15	***************
76.4	18	******************
76.5	21	*********************
76.6	12	************
76.7	8	********
76.8	9	*********
76.9	2	**
77.1	3	***
77.2	4	****
77.3	1	*
77.8	1	*

Boneless Beef Bead Boxes (M, N, O)

Boneless beef - 1 pound samples - 9,000 total pounds

Code Name	Key	Original Data	Contains	Color Key
M	Good	9,000 total 1 critical 7 major 100 minor	1 - critical defect 7 - major defects 100 - minor defects 8,892 - no defects 9,000 total	black blue yellow white
N	Minimum acceptable	9,000 total 3 critical 22 major 300 minor	3 - critical defects 19 - major defects 297 - minor defects 3 - minor-major defects 8,678 - no defects 9,000 total	black blue yellow pink white
O	Slightly unacceptable	9,000 total 6 critical 44 major 600 minor	4 - critical defects 36 - major 592 - minor 7 - minor-major 1 - minor-critical 1 - major-critical 8,359 - no defects 9,000 total	black blue yellow pink green purple white

PROBABILITY CALCULATION OF A USDA BONELESS BEEF PROBLEM

Composition of the total 9,000 pound lot. (Code name O)

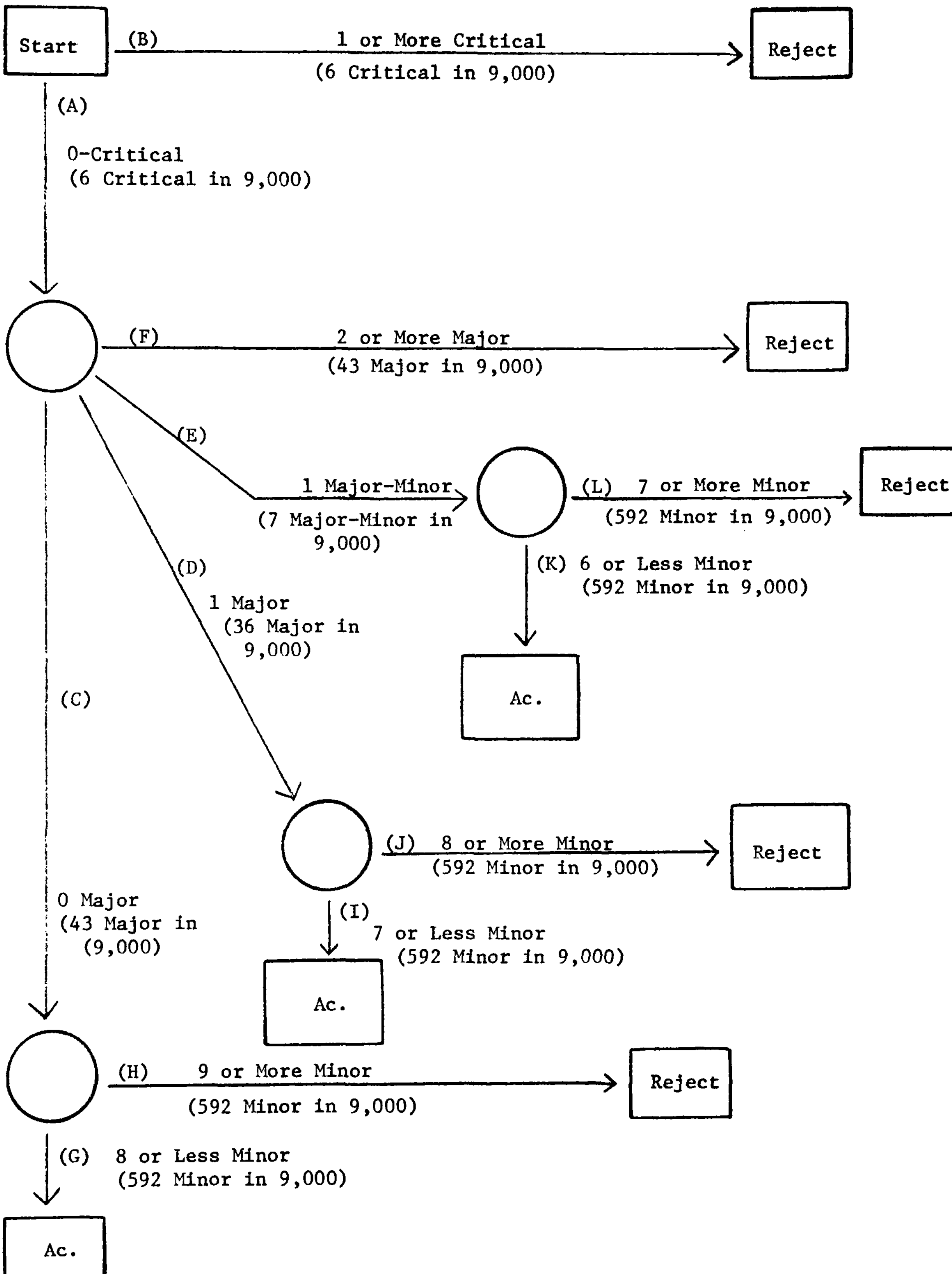
Condition of each pound	Number of pounds
Good	8,359
Critical	4
Major	36
Minor	592
Major-Minor	7
Critical-Minor	1
Critical-Major	1
	9,000 pounds total

Using the following Sampling Plan, what is the probability of accepting this lot.

SAMPLING PLAN

Pounds Sampled	Critical		Major		Total	
	Ac	Re	Ac	Re	Ac	Re
144	0	1	1	2	8	9

The following probability tree shows the routes that can be followed for accepting or rejecting this lot.


Start
(B)
1 or More Critical
(6 Critical in 9,000)
Reject
(A)
0-Critical
(6 Critical in 9,000)
(F)
2 or More Major
(43 Major in 9,000)
Reject
(E)
1 Major-Minor
(7 Major-Minor in 9,000)
(L) 7 or More Minor
(592 Minor in 9,000)
Reject
(K) 6 or Less Minor
(592 Minor in 9,000)
Ac.
(D)
1 Major
(36 Major in 9,000)
(C)
(J) 8 or More Minor
(592 Minor in 9,000)
Reject
(I)
7 or Less Minor
(592 Minor in 9,000)
Ac.
0 Major
(43 Major in (9,000)
(H) 9 or More Minor
(592 Minor in 9,000)
Reject
(G) 8 or Less Minor
(592 Minor in 9,000)
Ac.

Probability in Sampling one Pound.

$$\text{Critical total} = \frac{6}{9000} = .00067$$

$$\text{Major + Major-Minor} = \frac{43}{9000} = .00478$$

$$\text{Major only} = \frac{36}{9000} = .00400$$

$$\text{Major-Minor} = \frac{7}{9000} = .00078$$

$$\text{Minor only} = \frac{592}{9000} = .06578$$

Formulas that can be used to calculate the probability of each branch point.

$$\underbrace{\underbrace{\underbrace{(1-P)^n}_{\text{Probability of 0 occurring}} + \underbrace{\binom{n}{1} P^1(1-P)^{n-1}}_{\text{Probability of only 1 occurring}} + \underbrace{\binom{n}{2} P^2(1-P)^{n-2}}_{\text{Probability of only 2 occurring}}}_{\text{Probability of 0 or 1 or 2 occurring}} + \cdots + \underbrace{\binom{n}{c} P^c(1-P)^{n-c}}_{\text{Probability of only c occurring}}}_{\text{Probability of 0 through c occurring}}$$

P = Individual probability of that branch

n = Sample size 144 in this case

c = Number of samples with the condition specified

Calculation of the Probability of the Various routes (Represented by letters on the previous probability tree).

<u>0-Critical</u>

$$P_A = \left(1- \frac{6}{9000}\right)^{144} = .9080$$

<u>1 or More Critical</u>

$$P_B = 1 - .9080 = .0920$$

<u>Check</u>

$$P_B = \binom{144}{1} (.00067)^1(1-.00067)^{143} + \binom{144}{2} (.00067)^2(.99933)^{142} +$$

$$\binom{144}{3} (.00067)^3(.99933)^{141} + \binom{144}{4} (.00067)^4(.99933)^{140} +$$

$$\binom{144}{5} (.00067)^5(.99933)^{139} + \binom{144}{6} (.00067)^6(.99933)^{138} =$$

$$8.766244812 \times 10^{-2} + 4.202285108 \times 10^{-3} +$$

$$1.333578181 \times 10^{-4} + 3.151689902 \times 10^{-6} +$$

$$5.916534334 \times 10^{-8} + 9.189604411 \times 10^{-10} =$$

$$9.200124213 \times 10^{-2} = .0920$$

<u>0-Major</u>

$$P_c = \left(1 - \frac{43}{9000}\right)^{144} = .5017$$

1 Major

$$P_D = 144\ (.004)\ (1-.004)^{143} = .3247$$

1 Major Minor

$$P_E = 144\ (.00078)\ (1-.00078)^{143} = .1005$$

2 or More Major

$$P_F = 1-P_C-P_D-P_E = 1-.5015 -.3247 -.1005 = .0733$$

6, 7 and 8 or Less Minor

Probability of this number of Minor	Calculations	Results	Accumulative Totals
0	$(1-.06578)^{144} =$	$5.554936722 \times 10^{-5}$	
1	$(144)(.06578)(.93422)^{143}$	$5.632306973 \times 10^{-4}$	
2	$\binom{144}{2}(.6578)^2(.93422)^{142}$	$2.835548417 \times 10^{-3}$	
3	$\binom{144}{3}(.06578)^3(.93422)^{141}$	$9.450371158 \times 10^{-3}$	
4	$\binom{144}{4}(.06578)^4(.93422)^{140}$	$2.345593208 \times 10^{-2}$	
5	$\binom{144}{5}(.06578)^5(.93422)^{139}$	$4.6244004578 \times 10^{-2}$	
6	$\binom{144}{6}(.06578)^6(.93422)^{138}$	$7.543340187 \times 10^{-2}$	$1.580380381 \times 10^{-1}$
7	$\binom{144}{7}(.06578)^7(.93422)^{137}$	$1.047103108 \times 10^{-1}$	$2.627483489 \times 10^{-1}$
8	$\binom{144}{8}(.06578)^8(.93422)^{136}$	$1.262596954 \times 10^{-1}$	$3.890080443 \times 10^{-1}$

8 or Less Minor

$$P_G = .3890$$

9 or More Minor

$$P_H = 1-.3890 = .6110$$

7 or Less Minor

$$P_I = .2627$$

8 or More Minor

$$P_J = 1-.2627 = .7373$$

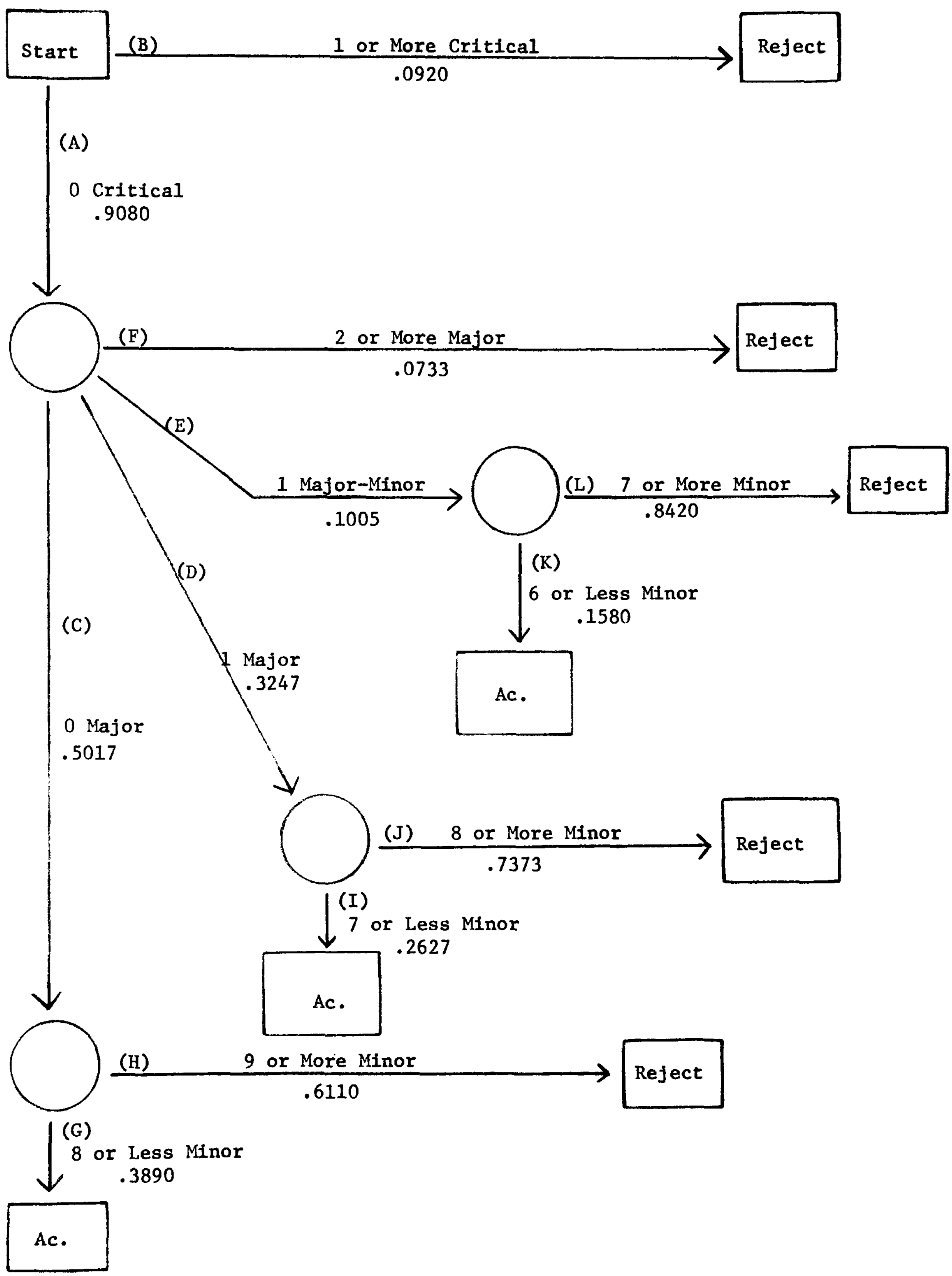
6 or Less Minor

$$P_K = .1580$$

7 or More Minor

$$P_L = 1-.1580 = .8420$$

Recording the probability on each route gives the following.


Start
(B) 1 or More Critical
.0920
Reject
(A)
0 Critical
.9080
(F) 2 or More Major
.0733
Reject
(E)
1 Major-Minor
.1005
(L) 7 or More Minor
.8420
Reject
(K)
6 or Less Minor
.1580
Ac.
(D)
1 Major
.3247
(C)
0 Major
.5017
(J) 8 or More Minor
.7373
Reject
(I)
7 or Less Minor
.2627
Ac.
(H) 9 or More Minor
.6110
Reject
(G)
8 or Less Minor
.3890
Ac.

Probability of Acceptance

$$P_1 = (A)(C)(G)$$

$$P_2 = (A)(D)(I)$$

$$P_3 = (A)(E)(K)$$

$$P_T = P_1 + P_2 + P_3$$

$$P_1 = (.9080)(.5017)(.3890) = .1772$$

$$P_2 = (.9080)(.3247)(.2627) = .0775$$

$$P_3 = (.9080)(.1005)(.1580) = \underline{.0144}$$

$$P_T = .2691 \text{ Acceptance}$$

Probability of Rejection

$$P_4 = (B)$$

$$P_5 = (A)(F)$$

$$P_6 = (A)(E)(L)$$

$$P_7 = (A)(D)(J)$$

$$P_8 = (A)(C)(H)$$

$$P_T = P_4 + P_5 + P_6 + P_7 + P_8$$

$$P_4 = (.0920) = .0920$$

$$P_5 = (.9080)(.0733) = .0666$$

$$P_6 = (.9080)(.1005)(.8420) = .0768$$

$$P_7 = (.9080)(.3247)(.7373) = .2174$$

$$P_8 = (.9080)(.5017)(.6110) = \underline{.2783}$$

$$P_T = .7311 \text{ Rejection}$$

Explanation of I_2 Values:

	For 99% level		For 95% level	
UCL_X and LCL_X =	$\bar{\bar{X}} \pm 3s$	or	$\bar{\bar{X}} \pm 2s$	
		$s = \frac{\bar{R}}{d^2}$		
=	$\bar{\bar{X}} \pm 3 \frac{\bar{R}}{d_2}$	or	$\bar{\bar{X}} \pm 2 \frac{\bar{R}}{d_2}$	
=	$\bar{X} \pm \frac{3}{d_2}(\bar{R})$	or	$\bar{\bar{X}} \pm \frac{2}{d_2} (\bar{R})$	
		For 99% level $I_2 = \frac{3}{d_2}$		For 95% level $I_2 = \frac{2}{d_2}$
=	$\bar{\bar{X}} \pm I_2\bar{R}$	or	$\bar{\bar{X}} \pm I_2\bar{R}$	

Explanation of A_2 Values:

$UCL_{\bar{X}}$ and $LCL_{\bar{X}}$ =	$\bar{\bar{X}} \pm 3s_{\bar{X}}$	or	$\bar{\bar{X}} \pm 2s_{\bar{X}}$
		$s_{\bar{X}} = \frac{s}{\sqrt{\text{n per group}}}$	
=	$\bar{\bar{X}} \pm 3 \frac{s}{\sqrt{n}}$	or	$\bar{\bar{X}} \pm 2 \frac{s}{\sqrt{n}}$
=	$\bar{\bar{X}} \pm \frac{3}{\sqrt{n}} (s)$	or	$\bar{\bar{X}} \pm \frac{2}{\sqrt{n}} (s)$
		$s = \frac{\bar{R}}{d_2}$	
=	$\bar{\bar{X}} \pm \frac{3}{\sqrt{n}} \cdot \frac{\bar{R}}{d_2}$	or	$\bar{\bar{X}} \pm \frac{2}{\sqrt{n}} \cdot \frac{\bar{R}}{d_2}$

(continued)

Explanation of A_2 Values: (continued)

	For 99% level		For 95% level	
$UCL_{\bar{X}}$ and $LCL_{\bar{X}}$	$= \bar{\bar{X}} \pm \left[\frac{3}{\sqrt{n}\,(d_2)}\right] \bar{R}$	or	$\bar{\bar{X}} \pm \frac{2}{\sqrt{n}\,(d_2)}\ \bar{R}$	
		For 99% level $A_2 = \frac{3}{\sqrt{n}\,(d_2)}$		For 95% level $A_2 = \frac{2}{\sqrt{n}\,(d_2)}$
	$= \bar{\bar{X}} \pm A_2\ \bar{R}$	or	$\bar{\bar{X}} \pm A_2\ \bar{R}$	
Explanation for D_3 and D_4 UCL_R and LCL_R	$= \bar{R} \pm 3s_R$	or	$\bar{R} \pm 2s_R$	
		$s_R = (s_w)s$		
	$= \bar{R} \pm 3(s_w)(s)$	or	$\bar{R} \pm 2(s_w)(s)$	
		$s = \frac{\bar{R}}{d^2}$		
	$= \bar{R} \pm 3(s_w)\ \frac{\bar{R}}{d_2}$	or	$\bar{R} \pm 2(s_w)\ \frac{\bar{R}}{d_2}$	
	$= \bar{R} \pm \left[3\ \frac{s_w}{d_2}\right] \bar{R}$	or	$\bar{R} \pm \left[2\ \frac{s_w}{d_2}\right] \bar{R}$	
	$= \bar{R}\ \left[1 \pm 3\ \frac{s_w}{d_2}\right]$	or	$\bar{R}\ \left[1 \pm 2\ \frac{s_w}{d_2}\right]$	
		For 99% level $D_4 = 1 + \frac{3s_w}{d_2}$		For 95% level $D_4 = 1 + \frac{2s_w}{d_2}$
		$D_3 = 1 - \frac{3s_w}{d_2}$		$D_3 = 1 - \frac{2s_w}{d_2}$

$UCL_R = D_4\ \bar{R}$

$LCL_R = D_3\ \bar{R}$

Therefore

$D_3 = 2 - D_4$